WHITE PAPER

INTELLIGENT BUSINESS STRATEGIES

The Value of Data Fabric in a Data-Driven Enterprise

By Mike Ferguson
Intelligent Business Strategies
June 2022

Research sponsored by:
StreamSets™

Table of Contents

Introduction ... 3

The hybrid multi-cloud operating model now supporting business operations ... 3

Challenges Caused By Todays Set-Up And Demands For Data ... 4

The complexity of a distributed data landscape ... 4

Fractured data governance ... 4

Multiple siloed analytical systems ... 4

A siloed approach to data integration ... 5

Requirements For Producing Data In A Distributed Data Landscape ... 7

Support for federated organisational setup ... 7

The need to produce reusable data products ... 8

Common enterprise data fabric software ... 8

Cloud ... 8
Connectivity ... 8
Data Catalog ... 9
Collaborative Development and Development Productivity ... 9
Data Capture and Processing ... 10
Resilience ... 10
Extensibility ... 10
Data Marketplace ... 11
Integration With Other Tools and Applications ... 11
Governance ... 11

Implementing Enterprise Data Fabric using StreamSets ... 12

The StreamSets DataOps Platform ... 12

StreamSets Control Hub ... 12
StreamSets Data Collector ... 13
StreamSets Transformer for Spark ... 13
StreamSets Transformer for Snowflake ... 13

StreamSets DataOps Platform Deployment Options ... 14

StreamSets DataOps Platform Capabilities ... 14

StreamSets DataOps Platform Technology Integrations ... 15

Conclusions ... 16

INTRODUCTION

Most executives now see data and analytics as essential in every-day business

In almost every company today, executives have already ‘bought in’ to the strategic importance of data and analytics in every-day business. It is no longer a ‘nice to have’ capability as it was in past years. It is an essential capability. So much so that the need to prepare and analyse data is no longer a central IT function done on behalf of business departments. It has become a requirement in every part of a business from marketing and finance who have always sponsored analytical projects, to business operations right through to back-office functions like procurement and human resources. Today, analytical systems have become the ‘brains’ of modern business with business intelligence, machine learning (ML) models and artificial intelligence being produced right across the business and ‘wired’ into every application and business process.

Analytical systems have become the ‘brains’ of modern business and are in use enterprise wide

New data sources are in demand as companies look to gain deeper insights beyond what they already know

Not surprisingly, such unprecedented demand for insights and AI from across the business has seen major growth in demand for data. However, the data required is not just from traditional transaction processing systems but from new data sources as companies look to gain deeper insights beyond what they already know. This new data is coming from within and from outside the enterprise. Everything from inbound email, customer chat, social network opinions, on-line clickstream data, IoT sensor data, external open government data, financial markets data and weather data are all now being added to traditional data sources and analysed on multiple different analytical systems.

THE HYBRID MULTI-CLOUD OPERATING MODEL NOW SUPPORTING BUSINESS OPERATIONS

Perhaps the major difference today from past years is that most companies are now operating on a distributed, multi-cloud, hybrid computing environment that spans on-premises, multiple clouds, and the edge. This is shown in Figure 1. They are also analysing data in this environment with different types of analyses happening on-premises, in one or more clouds and also at the edge.

Data is being captured and stored in different types of data store across a hybrid multi-cloud distributed computing environment

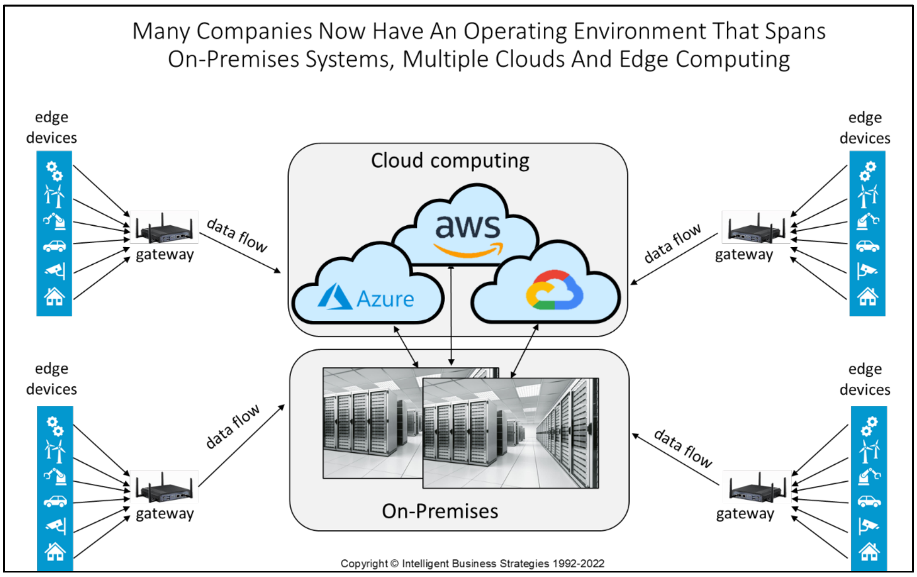

Many Companies Now Have An Operating Environment That Spans
On-Premises Systems, Multiple Clouds And Edge Computing
edge devices
edge devices
edge devices
edge devices
gateway
gateway
gateway
gateway
data flow
data flow
data flow
data flow
Cloud computing
aws
Azure
On-Premises
Copyright © Intelligent Business Strategies 1992-2022

Figure 1

Challenges Caused By Todays Set-Up And Demands For Data

There are a number of challenges that come from operating on this kind of hybrid, multi-cloud distributed computing set-up. These are discussed below.

The complexity of a distributed data landscape

Data complexity has increased as a result of operating on a hybrid, multi-cloud distributed computing environment

The first is data complexity. We now have data being generated, streamed, created, captured, and persisted in different types of data store at the edge, in SaaS applications, in multiple clouds and on-premises. Data may be streamed into scalable messaging systems such as Kafka, AWS Kinesis, Google Pub/Sub. It may be stored in edge databases, data lakes, Hadoop systems, No-SQL databases and relational databases across this hybrid, multi-cloud distributed computing environment. Therefore, what we have is a complex distributed data landscape with more and more valuable data sources emerging and being ingested into the enterprise.

As a result, data can be anywhere making it more difficult for business users to find

Given this complexity, it is not surprising that business users are struggling to find the data they need. After all, it could be anywhere across a broad range of streaming data sources and multiple different types of data store. Not only that, but the number of copies of data that can exist across this hybrid, multi-cloud distributed computing environment adds to the challenge of finding the right data. And with so many different types of data store, people may be uncertain as to which copy is the most up to date.

Fractured data governance

Governing data has become more difficult and fractured across an increasingly complex distributed data estate

Another major challenge is that governance of data across this increasingly complex distributed data estate is fractured. Managing data privacy, access security, data sharing, data retention, and data quality across so many data stores and SaaS applications is siloed and labour intensive. It is reliant on an increasing number of administrators to implement the same data governance policies repeatedly across different data stores using many different administrative tools. Furthermore, as companies move more data to the cloud there is concern that they can no longer control their governance practice. This is primarily because with the cloud, and with SaaS applications, administrators of systems may be outside the organisation. The simple fact is that with more data sources, data lakes containing thousands if not millions of files, and databases now residing outside the corporate firewall there is concern that fractured data governance increases the risk of non-compliance, accidental oversharing of data, data breach, and other governance related issues that could cause harm to the organisation.

Multiple siloed analytical systems

In addition, all this data has spawned a range of analytical workloads many of which are running on different types of analytical system. This includes data warehouses, data lakes (e.g., Cloud storage, Hadoop), in use by data scientists, lake houses, NoSQL graph databases. It also includes streaming analytics happening on data streaming platforms like Kafka, AWS Kinesis, or Google Pub/Sub. Today, most companies have some combination of these. However, they are almost always running as silos as shown in Figure 2.

Many companies now have a number of independently operated, siloed analytical systems

These systems could be operating on-premises, in the cloud and at the edge but they are not integrated

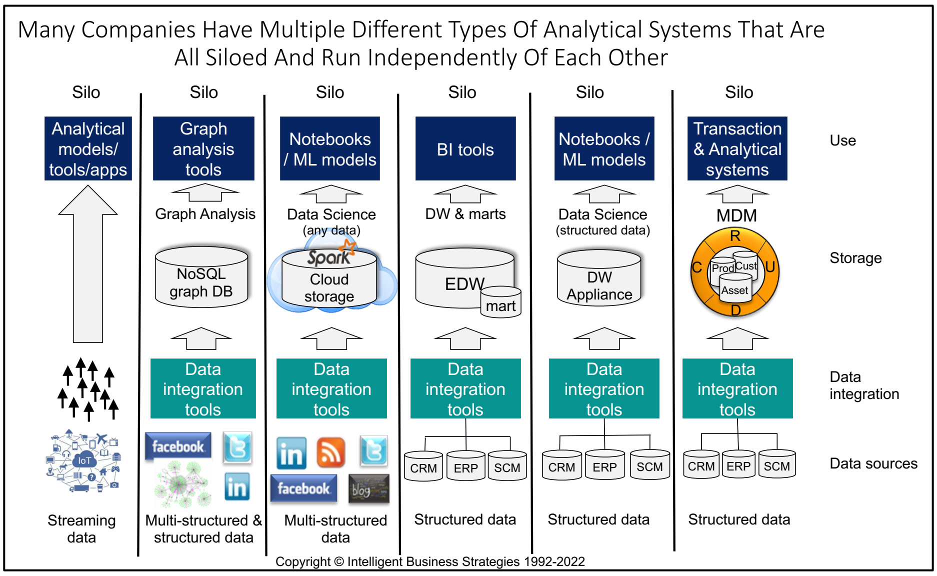


Figure 2

With this kind of approach work is being repeated across analytical system silos.

A SILOED APPROACH TO DATA INTEGRATION

Looking at Figure 3 the one area that stands out is data integration where different data integration tools are often used in each different analytical system.

Work such as data integration is being repeated across different analytical systems often using different tools

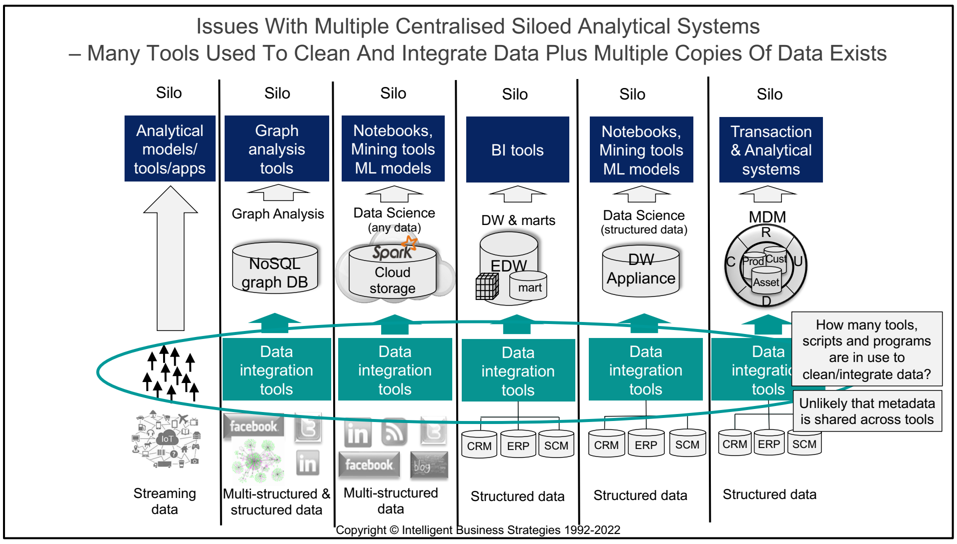

Issues With Multiple Centralised Siloed Analytical Systems
– Many Tools Used To Clean And Integrate Data Plus Multiple Copies Of Data Exists
Silo
Silo
Silo
Silo
Silo
Silo
Analytical models/ tools/apps
Graph analysis tools
Notebooks, Mining tools ML models
BI tools
Notebooks, Mining tools ML models
Transaction & Analytical systems
Graph Analysis
Data Science (any data)
DW & marts
Data Science (structured data)
MDM
NoSQL graph DB
Spark
Cloud storage
EDW
mart
DW Appliance
R
C
Prod
Cust
U
Asset
D
Data integration tools
Data integration tools
Data integration tools
Data integration tools
Data integration tools
How many tools, scripts and programs are in use to clean/integrate data?
Unlikely that metadata is shared across tools
IoT
facebook
in
blog
CRM
ERP
SCM
CRM
ERP
SCM
CRM
ERP
SCM
Streaming data
Multi-structured & structured data
Multi-structured data
Structured data
Structured data
Structured data
Copyright © Intelligent Business Strategies 1992-2022

Figure 3

Speed of development is slower because very little is being shared and reused across multiple tools

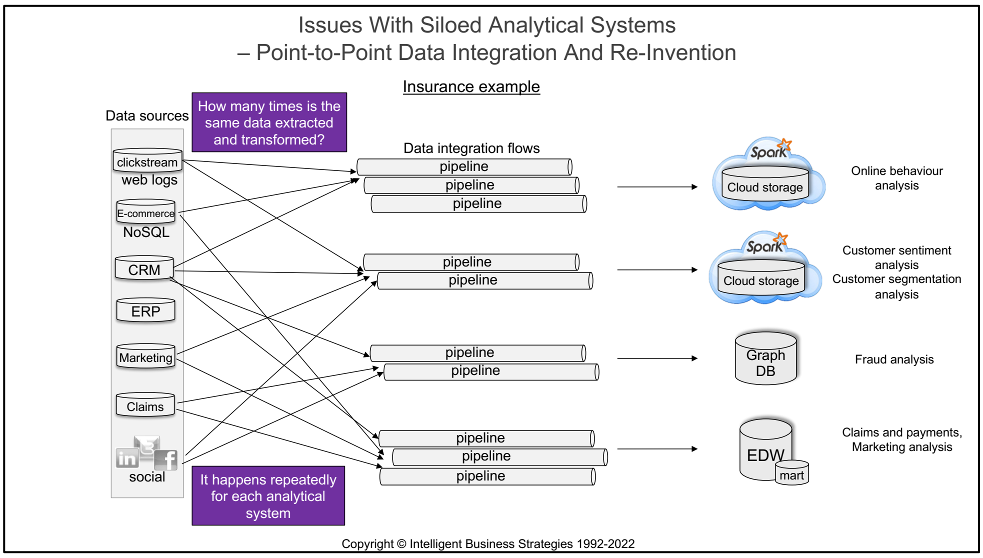
The tool used to capture, clean, transform and integrate data to populate a data warehouse for example, is unlikely to be the same as that used to do the same for a graph database or for streaming analytics. A good question to ask is “how many tools, scripts, programming languages (e.g., databases stored procedure languages, Python, Scala, R, Java...) are in use to do this in your organisation?”.

Reinvention and multiple tools have led to the cost of data integration being higher than it should be

The impact of this is significant. It slows speed of development because nothing is shared across silos. If you add in the number of business analysts also doing self-service data preparation using even more tools, then it is easy to see that the cost of data integration is way too high. The problem is that most companies have never attempted to quantify this and so can’t easily put a figure on it. But, as new data sources and data stores continue to emerge, the cost of building and maintaining data integration jobs could easily spiral out of control.

In addition, working to clean, transform and integrate data in multiple independent analytical siloes is going to lead to much more re-invention rather than sharing and re-use. How many times, is the same data extracted, cleaned, transformed, and integrated for use in each different analytical system or for input into different machine learning algorithms in data science? An insurance example of this is shown in Figure 4 where different data integration pipelines are built for each different analytical system, in many cases repeatedly taking the same data from the same data sources.

Point-to-point data integration pipelines built using different tools frequently capture and transform the same data for the same sources


Issues With Siloed Analytical Systems
– Point-to-Point Data Integration And Re-Invention
Insurance example
Data sources
How many times is the same data extracted and transformed?
clickstream
web logs
E-commerce
NoSQL
CRM
ERP
Marketing
Claims
social
Data integration flows
pipeline
pipeline
pipeline
pipeline
pipeline
pipeline
pipeline
pipeline
pipeline
pipeline
It happens repeatedly for each analytical system
Spark
Cloud storage
Online behaviour analysis
Spark
Cloud storage
Customer sentiment analysis
Customer segmentation analysis
Graph DB
Fraud analysis
EDW
mart
Claims and payments, Marketing analysis

Copyright © Intelligent Business Strategies 1992-2022

Figure 4

Inability to share metadata across multiple tools leads to reinvention as people are unaware of what has already been created

Much of this is caused by different stand-alone data integration tools not being integrated. So, metadata is fractured across tools with little to no chance of sharing it. Therefore, what is invented in one tool is often unknown to another. In addition, there is no universal standard to exchange metadata between tools and if people write code (e.g., stored procedures or Python) then lineage is unavailable making it difficult to know what a data engineer did to prepare data.

Skillsets may also be thinly spread across many tools

It is also likely that skillsets may be too thinly spread across many tools which increases the risk of no one knowing how to maintain pipelines if data engineers should ever choose to leave.

Many pipelines have become brittle over time as change requests accelerate due to more data becoming available

This increases the cost of maintenance

Finally, there is a high chance that many of the pipelines built to clean, transform and integrate data are brittle because dependent data flows have been added into pipelines over the years as change requests have occurred which makes them more complex and more prone to failure. This is increasingly a problem because as more data sources become available, the request for change to add new data accelerates. The growth in new and changing data sources means that source schemas are changing. In many cases (e.g., IoT), this is happening without notice and so if a data integration pipeline can’t accommodate this, it is highly likely to break. Given the aforementioned multitude of stand-alone data integration tools and pipeline ‘add ons’, just think of the cost to maintain brittle pipelines across all those tools. Not only that, but if it is just a centralised team of people doing this on behalf of the business, they are highly likely to become a bottleneck that can’t keep pace with business demands.

There has to be a better way. In order to understand that it is first necessary to understand the requirements for producing trusted, compliant, re-usable data in a distributed data landscape.

REQUIREMENTS FOR PRODUCING DATA IN A DISTRIBUTED DATA LANDSCAPE

The following set of requirements act as a guide to what is needed to produce trusted, compliant, re-usable data in a distributed data landscape.

SUPPORT FOR FEDERATED ORGANISATIONAL SETUP

Companies need to industrialise the process of producing high quality, compliant data that can be easily shared across the enterprise

The first thing to recognise is that companies need to 'industrialise' the process of producing high quality, compliant data that can easily be shared around the organisation. That means scaling up the number of people doing data engineering but doing this in an organised way that can incrementally shorten time to value. There needs to be recognition that we are now in a world of data producers and data consumers and also a need to co-ordinate this activity to avoid the pitfalls described earlier.

Therefore, rather than just decentralised teams doing their own thing, a federated organisational set-up is needed with:

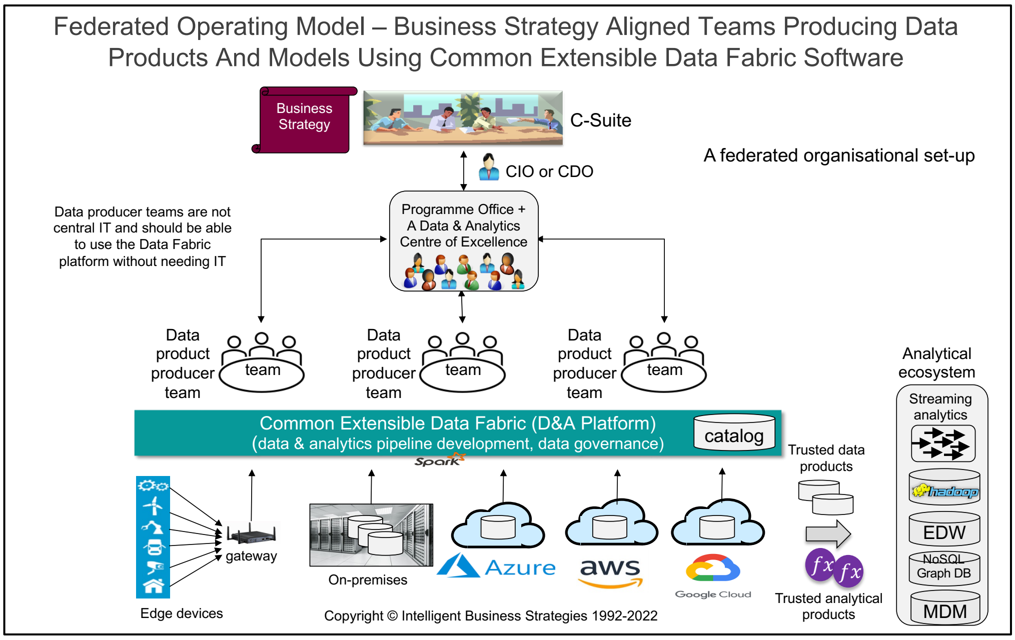
It is also important that companies organise to succeed

- An executive accountable for ensuring that it is as easy as possible for data producers to create high quality, compliant, sharable data products
- A centralised program office to co-ordinate projects so that everyone knows who is producing what data and for what business purpose
- A data and analytics centre of excellence with experts available to support and embed in data producer teams around the organisation to help them quickly and easily produce data. These teams can include citizen data engineers, business domain subject matter experts (SMEs) and embedded professional data engineering experts.

This is shown in Figure 5.

A federated operating model with a centralised program office and a centre of excellence will help co-ordinate activity across multiple teams

Allowing multiple teams to share a common data fabric enables metadata sharing, prevents reinvention, improves productivity, and accelerates production of sharable data


Federated Operating Model – Business Strategy Aligned Teams Producing Data Products And Models Using Common Extensible Data Fabric Software

Business Strategy
C-Suite
A federated organisational set-up
CIO or CDO
Programme Office + A Data & Analytics Centre of Excellence
Data producer teams are not central IT and should be able to use the Data Fabric platform without needing IT
Data product producer team
team
Data product producer team
team
Data product producer team
team
Analytical ecosystem
Streaming analytics
hadoop
EDW
NoSQL Graph DB
MDM
Common Extensible Data Fabric (D&A Platform)
(data & analytics pipeline development, data governance)
catalog
Spark
Trusted data products
Trusted analytical products
gateway
Edge devices
On-premises
Azure
aws
Google Cloud

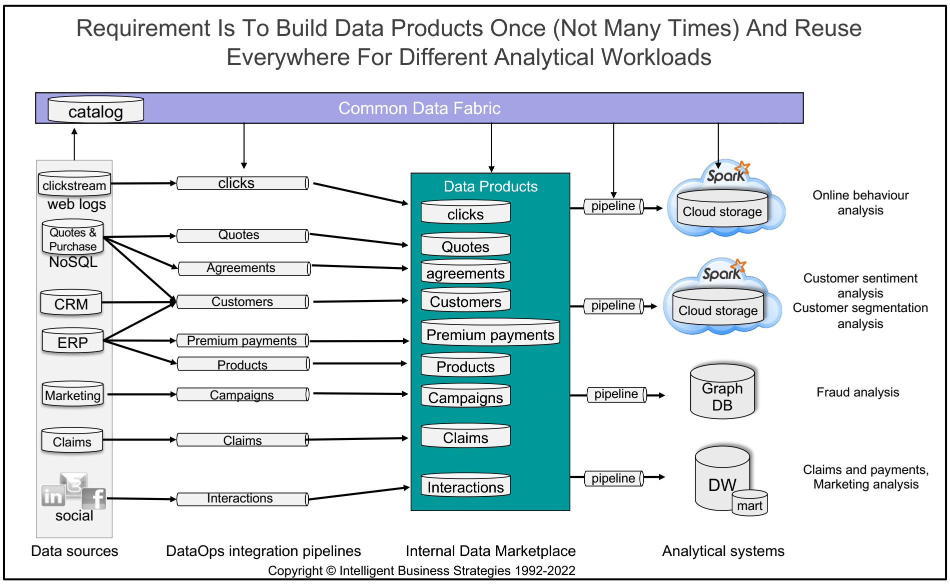
Copyright © Intelligent Business Strategies 1992-2022

Figure 5

THE NEED TO PRODUCE REUSABLE DATA PRODUCTS

Staying with Figure 5, rather than data producer teams creating data pipelines to populate a schema associated with a single analytical system, they instead need to produce high quality, compliant data products that can be reused across multiple analytical systems and workloads. Figure 6 shows that idea behind creating data products. It means that data entities can be built separately and then the required data products can be selected and 'assembled' to support a specific analytical workload whether that be a data warehouse, a graph database or to provide features as input to an algorithm during machine learning model development in data science.

Data producers need to focus on building reusable data products rather than building pipelines just to populate a single system or to provide features for input into a single machine learning model



Requirement Is To Build Data Products Once (Not Many Times) And Reuse Everywhere For Different Analytical Workloads

Copyright © Intelligent Business Strategies 1992-2022

Figure 6

COMMON ENTERPRISE DATA FABRIC SOFTWARE

In addition, rather than every team of data producers using different tools to create data products and run into the problems described earlier, it should be possible to enable multiple teams of data producers to utilise a common data fabric software platform that supports the following capabilities:

Multiple teams of data producers sharing a cloud-based common data fabric software platform is likely to accelerate production of reusable data products

Cloud

- Cloud deployment to enable one or more data producer teams to get started quickly using a subscription-based pay-as-you-use approach

Connectivity

- Connectivity to as many data sources and targets as possible across a distributed data estate. This would include data sources such as:
  - Streaming data services such as scalable data messaging software like Kafka, AWS Kinesis, Google Pub/Sub, Azure Event Hubs
  - Cloud data stores and applications on multiple different clouds including cloud storage, cloud based no-SQL databases, cloud-based relational databases, and cloud-based Hadoop systems
  - On-premises data streaming, data stores and applications including messaging queuing and service bus software, flat files, on-premises no-SQL databases, relational databases, and Hadoop systems

Data fabric software needs to support a broad range of data connectors

  - SaaS applications running in application vendor data centres with connectivity via APIs or access to underlying data stores
  - Edge devices and edge databases

Note that edge-based connectivity is important so that data producers can build pipelines to capture, process and produce data products from real-time data as it is generated (e.g., by IoT devices) at the edge but also to provide the option to run the pipelines at the edge local to where the data originates.

Data Catalog

- A data catalog or integration with a third-party data catalog to enable:
  - Automatic discovery of data, schema, and data relationships within and across multiple data sources in a distributed data estate
  - Automatic mapping of physical data names to business terms in a business glossary to understand the meaning of data and provide support for semantic classification
  - Automatic classification of sensitive data types to know what data to protect with respect to data privacy and access security

A data catalog is needed to help data producers quickly understand what data is available and where sensitive data is located

This capability enables data producers to quickly understand what raw data is available to them to create data products and where sensitive data is so they can anonymise it to keep it protected in line with compliance obligations.

Collaborative Development and Development Productivity

Data fabric software needs to support collaborative development within and across teams

- Multiple projects running simultaneously on the same platform to organise, manage and govern the development of multiple different data products by several different teams
- Role-based user interfaces to enable professional and citizen data engineers to jointly develop component-based pipelines to produce data products. In this way citizen and expert data engineers can work on different components to clean, transform and integrate data and then orchestrate these components together into a pipeline that produces a required data product.

Cater for professional and citizen data producers

- The ability to easily share metadata, across multiple data engineers within the same team and multiple teams. This would include sharing metadata about data sources, business glossary, data lineage, and data products that have been produced. Ideally this should be made possible by enabling multiple teams to have shared access to a searchable data catalog from within the data fabric. This should enable the ability to see relationships in data and also to drill down into the components of pipelines to see the transformations on data during data product development
- The ability to easily share pipeline components and templates across multiple data engineers within the same team and multiple teams
- Augmentation using ML to accelerate pipeline development and accelerate the creation of data products. This includes using augmentation to automatically:
  - Infer data source schema
  - Identify data quality issues in the data e.g., duplicates, missing data
  - Recommend transformations and cleansing to speed up development

Project orientation, shared metadata, reusable transformations, pre-built templates, and augmentation all help shorten development time

- Recommend mappings to speed up development of data integration

- Ability to see and dynamically infer schema at any point in a pipeline to easily test and debug transformations during pipeline development

DataOps for collaborative development, automated testing, automated deployment, and version control

- DataOps for collaborative development and version control with support for:
  - Branch, pull requests and merge for continuous development and continuous delivery (CI/CD)
  - Integration with universal code repositories (e.g., Git, GitHub, Bitbucket etc.) for transformation component and pipeline versioning
  - Automated testing and configuration management
  - Automated deployment

Data Capture and Processing

Ability to handle real-time streaming data, change data capture and batch extraction

- Processing data in-motion and at rest at scale including support for:
  - Structured, semi-structured and unstructured data sources
  - Real-time data ingestion and change data capture (CDC)
  - Extensive pre-built functions for data cleansing, data transformation, data enrichment and data matching
  - Multi-target pipelines to enable data to be built once and consistently delivered to multiple analytical workloads
  - Event-driven, and batch-oriented execution of pipelines
  - Executing pipelines at scale using multi-threading and Apache Spark

Ability to scale to handle high velocity data and large data volumes

Resilience

- Automatic detection and management of unexpected and undocumented changes in data structure and semantics with minimal impact. This would include run-time automatic detection of schema change or field order change in a data source plus the ability to 'design in' processing to manage this should it occur. This is a critical capability to handle the increasing number of data changes. It removes the burden on data engineers, prevents pipeline failure in the event of such changes. This is particularly important in IoT data sources but could apply to any data source.

Ability to automatically detect and easily accommodate change to avoid pipeline failures and reduce the burden of maintenance

- Automatic detection and management of change to infrastructure e.g., change in source database from SQL server to Oracle while all other requirements remain the same. It should be possible to:
  - Create, track, and maintain a new version of a pipeline
  - Easily revert back to the older version
  - Visually compare old and new versions
  - Parameterise a pipeline to enable different instances of the pipeline to run to connect to different data sources
- Ability to restart pipelines at the point where they stopped

Extensibility

- Inclusion of custom code (e.g., a Jupyter notebook) in pipelines to handle complex transformations

Data fabric software needs to be extensible to handle complex transformations and to utilise machine learning models developed in other tools

- Invocation of user defined functions (UDFs) and remote web services in pipelines to integrate data transformations and analytical services developed in other tools
- Inclusion of machine learning models in pipelines during pipeline execution e.g., auto data matching, natural language processing
- Inclusion of pre-built and 3rd party cognitive services in pipelines e.g., voice-to-text conversion, sentiment analysis, language translation

Data Marketplace

A data marketplace to publish reusable data products and make them available for consumption

- A searchable data marketplace to govern the publishing, sharing and consumption of data products including support for:
  - Search and faceted search to help consumers quickly find data
  - Data quality scores associated with published data products
  - Data product version management
  - Consumer access to a business glossary describing the meaning of data available in each data product
  - Consumer access to full metadata lineage to understand how pipelines process data to create each data product
  - Definition of policies to control access security, privacy, sharing and retention of data products
  - Definition of data sharing agreements so terms and conditions associated with sharing data are accepted and audited before the data is shared. This enables organisations to track usage and ensure compliance with regulatory and legislative obligations

Integration With Other Tools and Applications

APIs to integrate with other tools so developers can use the platform to prepare data rather than code it all themselves

- A set of fully documented APIs to:
  - Enable software developers and data scientists to make use of the full range of data services on the data fabric platform
  - Facilitate integration with 3rd party software products

Governance

Ability to protect sensitive data to remain compliant with regulations and to monitor pipeline execution

- Integration with common enterprise services – e.g., Active Directory / LDAP
- Automatic detection, dynamic data masking and encryption of sensitive data during pipeline execution to ensure sensitive data is protected on ingest and during processing and to ensure the creation of compliant data products
- Provide continuous observability analytics of pipelines to enable continuous visibility and monitoring at every stage of execution.
- Built-in data governance through the ability to define policies in the data catalog and the ability to enforce them across multiple systems via connectors and/or APIs that can be used to integrate with other technologies

IMPLEMENTING ENTERPRISE DATA FABRIC USING STREAMSETS

Having discussed the requirements for producing data in a distributed data landscape, this section of the paper looks at how StreamSets steps up to meet those requirements. StreamSets is a SoftwareAG company founded in 2014 that has customers all over the world in many different vertical industries.

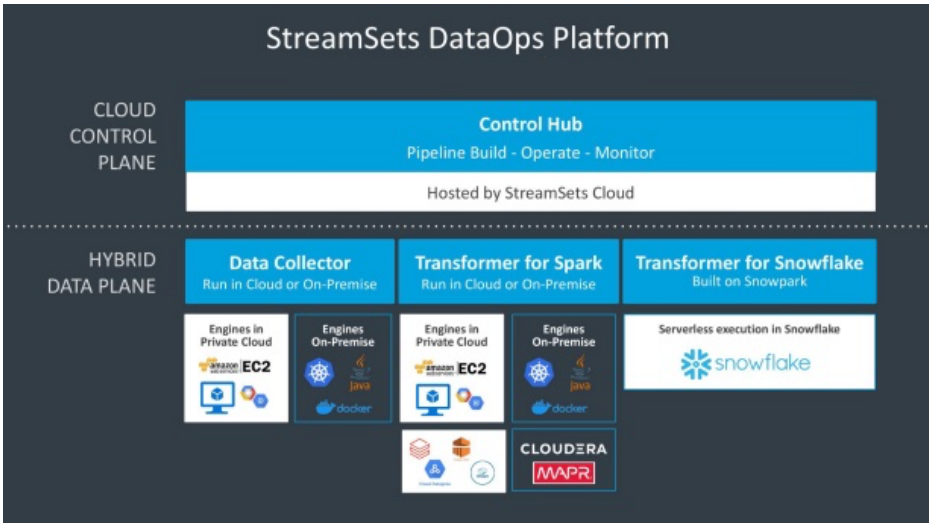
THE STREAMSETS DATAOPS PLATFORM

StreamSets' data fabric offering is the StreamSets DataOps Platform which allows you to design and build data pipelines across a hybrid, multi-cloud distributed data landscape that are resilient to unexpected changes in source system schema, semantics, and infrastructure. StreamSets DataOps Platform offers a broad range of data connectors and consists of a number of components shown in Figure 7. These are:

The StreamSets DataOps Platform consists of a number of components

- StreamSets Control Hub
- StreamSets Data Collector
- StreamSets Transformer for Spark
- StreamSets Transformer for Snowflake

Pipelines can be developed in the cloud and can run on-premises and on one or more clouds in a hybrid multi-cloud computing environment



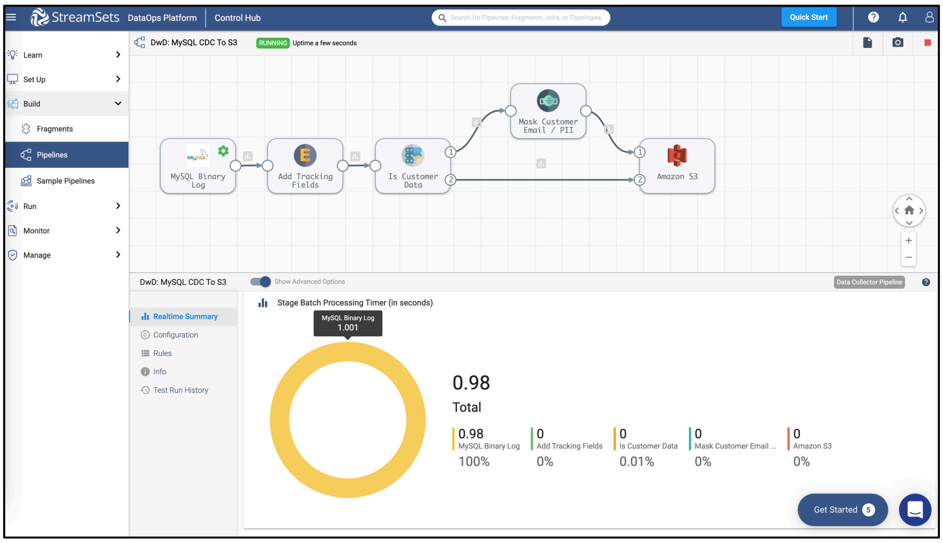
Source: StreamSets

Figure 7

StreamSets Control Hub

The StreamSets Control Hub provides a central web-based console to build, run and track pipelines

StreamSets Control Hub provides a centralised real-time, console that allows you to build, run and track hundreds of streaming, change data capture (CDC), batch ETL and ML data pipelines that can process data on-premises and on multiple clouds. Control Hub provides a web-based user interface (see Figure 8) to build and configure pipelines, preview data, review snapshots of data and monitor pipelines. This includes viewing real-time metrics about these jobs such as record counts, throughput, processing times etc.



Source: StreamSets

Figure 8

StreamSets Data Collector executes pipelines to capture and ingest data into data lake ingestion zones, lake houses, and data warehouse staging areas

StreamSets Data Collector

StreamSets Data Collector is an execution engine that works directly with StreamSets Control Hub. The purpose of a Data Collector is to execute pipelines to capture and ingest data into data lake ingestion zones, lake houses and data warehouse staging areas where it can be transformed to produce data products for use in multiple analytical workloads. It is also possible to use StreamSets Data Collector pipelines to collect live streaming data, process and analyse it in real-time without the need to store it first.

Data Collector agents allow pipelines to run close to the data where it can be captured, filtered, processed, and masked as it is collected

During execution of a data collection pipeline, data can be filtered, data types converted, fields renamed, re-ordered, split, merged, IP addresses geo-coded, sensitive data masked, etc., etc.

Pipelines can be multi-threaded and run on cluster configurations to process large volumes of data in parallel at scale. They can also be configured to be restartable to avoid re-processing large data volumes if the pipeline fails.

StreamSets DataOps Platform has two StreamSets Transformer execution engines

StreamSets Transformer for Spark

StreamSets Transformer for Spark uses Apache Spark to execute scalable ETL and machine learning pipelines on-premises and/or in the cloud

StreamSets Transformer for Spark is an execution engine within the StreamSets DataOps platform that allows users to execute pipelines on Apache Spark. This enables pipelines to perform ETL, stream processing (Spark Streaming) and machine learning operations by utilising the Spark Machine Learning library (MLlib). The drag and drop user interface means that data analysts, data scientists and ETL developers can fully utilise the power of Apache Spark without requiring a deep technical understanding of the platform.

StreamSets Transformer for Snowflake

StreamSets Transformer for Snowflake executes 'in-database' transformations natively within the cloud-based Snowflake DBMS

StreamSets Transformer for Snowflake is a StreamSets Engine that executes transformations natively within the Snowflake DBMS. Transformations are pushed down into Snowflake to run in a dedicated virtual warehouse compute cluster to avoid impacting query workloads. This means that data can stay in Snowflake (or in cloud storage pointed to by Snowflake external tables) and everything runs natively under the governance of the Snowflake platform. It also allows StreamSets to embrace Snowflake's Snowpark environment meaning that developers can extend the capabilities of StreamSets by coding their own custom transformations in various programming languages, create the transformations as user defined functions (UDFs) in Snowflake and then include them as transformations in a StreamSets pipeline.

StreamSets DataOps Platform Deployment Options

As for deployment, StreamSets Control Hub is hosted on StreamSets Cloud.

StreamSets DataOps Platform provides the flexibility needed to run pipelines on-premises and on multiple clouds

StreamSets Data Collectors can be installed to run anywhere on your corporate network where your data lives. Data Collectors can be automatically provisioned in Docker containers to run on Kubernetes clusters on-premises or on any cloud computing platform. Once installed StreamSets Data Collectors are then registered to work with StreamSets Control Hub.

StreamSets Transformer for Spark can be deployed on-premises (e.g., Spark running on a Cloudera's CDP Hadoop system) or Spark running as a service on Databricks, Google Cloud, Microsoft Azure, or AWS.

StreamSets Transformer for Snowflake obviously runs on Snowflake which itself may be running on Google Cloud, Microsoft Azure, or AWS.

StreamSets DataOps Platform Capabilities

StreamSets DataOps Platform meets many of the data fabric requirements defined earlier in the paper.

StreamSets provides data fabric software that can connect to a broad range of data sources

It supports an extensive set of pre-built connectors to a broad range of data sources including IoT devices, streaming data services, cloud storage, cloud DBMSs, lake houses, on-premises NoSQL and relational DBMSs, on-premises messaging systems, Hadoop, web services, files and SaaS applications.

Multiple teams can use the platform to simultaneously develop pipelines all of which are governed from a single console

Structured, semi-structured and unstructured data can be captured and transformed

The platform enables multiple teams of data producers to build data pipelines on the cloud in support of decentralised development of data products. StreamSets Control Hub can be used to coordinate this effort as it can see all pipelines. Data can be collected from sources across a distributed data estate using streaming data, change data capture (CDC) and batch extraction-based ingestion techniques. Data Collector pipelines can capture structured, semi-structured and unstructured data. Pipelines can be built to parse JSON data, machine generated data in log files and live streaming data being generated directly at the edge or flowing in on scalable messaging systems like Kafka. It is also possible to process social network data (e.g., Twitter).

Scalable execution engines provide scalability

This data can then be cleaned, transformed, enriched, and integrated using a rich set of pre-built transformations to produce high quality, compliant, re-usable data products for use in multiple analytical workloads. These transformations can be multi-threaded and run on-premises, in one or more clouds or a combination of these.

Pipelines can automatically deal with unexpected changes in schema, semantics, and changes in infrastructure

Furthermore, StreamSets pipelines are both extensible and can be designed to be resilient. Extensibility is available in both StreamSets Transformer for Spark and StreamSets Transformer for Snowflake as the platform supports the inclusion of custom code that can run in Spark or as UDFs in Snowflake. In terms of resilience, StreamSets isn't hardwired into the systems it connects and can infer schema at runtime. Therefore, pipelines can be designed to automatically accommodate unexpected changes in source system schema, semantics, and changes in infrastructure. This reduces the number of pipeline failures that would normally occur in such circumstances. You can also mask and encrypt sensitive data as it flows through a pipeline to protect it in line with compliance obligations and link multiple jobs together so data can traverse collector and transformation pipelines in an end-to-end orchestration.

CI/CD support enables collaborative development and rapid deployment

StreamSets Control Hub also enables collaborative development, automated testing, and deployment of DataOps pipelines by supporting continuous integration / continuous deployment (CI/CD). This is achieved by setting subscriptions in StreamSets Control Hub so that any change to a pipeline triggers a Jenkins job to automatically test and deploy the changed pipeline and upgrade it to the new version within the Control Hub if it passes the test.

Transforms, templates, and metadata can be shared across StreamSets pipeline developers to improve productivity and a

In addition, transforms, templates, and metadata can be shared across pipeline developers who are all using the StreamSets Control Hub.

Developers and data scientists can utilise the platform to work on their behalf via SDKs and APIs

StreamSets DataOps Platform also provides a programmatic way to implement pipelines using an SDK for Python. This enables data scientists and developers to use the power of the platform to engineer data rather than code it all themselves. Using the SDK, Python programmers can create, preview, and run a pipeline, manage job lifecycles, and orchestrate deployment. StreamSets also provide a Python based test framework for writing automated test cases with support for the pytest test automation capability and Docker packaging.

Data products can be produced and published as services in a Data Mesh

Once data products are produced it is also possible to publish them as services in a Data Mesh by using a pipeline to create an interface to each data product. It is also possible to create additional pipelines to consume data products and assemble the required data products to support different analytical workloads.

StreamSets DataOps Platform Technology Integrations

Pre-built integration with application middleware and IoT platforms broadens the number of data sources and makes it possible to build intelligent applications

In addition to its own capabilities, StreamSets DataOps Platform has pre-built integration with a number of other technologies to broaden the number of data sources and open up the platform to integrate with 3rd party data governance tools. Now that StreamSets is a SoftwareAG company, planned integration with SoftwareAG's application middleware and IoT platforms will enable pipelines to reach more application and IoT data sources and to deliver StreamSets ML pipeline predictions, recommendations, and insights back into applications and devices.

Pre-built integration with 3rd party data governance tools and cloud data platforms is also available

With respect to integration with data governance tools, StreamSets Data Collector can be configured to publish metadata about running pipelines to Collibra and Apache Atlas. You then use Collibra or Apache Atlas to explore the pipeline metadata, including viewing metadata lineage diagrams explaining how data was processed in pipelines.

There is also integration with Snowflake with StreamSets Transformer for Snowflake running 'in-database' transformation pipelines on virtual warehouse clusters within Snowflake.

Conclusions

Companies need an organised and co-ordinated approach to accelerate the development of reusable data products for use in analytics

If companies are serious about becoming data-driven they need to be able to scale up the number of teams around the enterprise to engineer data in order to produce reusable data products. However, it needs to be organised and co-ordinated via a central program office and a centre of excellence with experts helping to upskill these teams to help them become more productive. The objective is to build them once and reuse them everywhere.

This is much easier and less costly to do using a common data fabric platform rather than multiple stand-alone tools

Trying to do this using a wide range of different stand-alone data integration tools slows speed of development because nothing can be shared across tools. This approach is also likely to lead to more re-invention and the cost of data integration is likely to be much higher than necessary. Instead, it should be possible to enable multiple teams of data producers to utilise a common data fabric software platform, preferably on the cloud where metadata, templates and transformation components can be easily shared across multiple teams.

A common data fabric platform needs to support broad connectivity, provide access to a data catalog, and enable collaborative development and sharing

It should be possible to enable multiple teams of data producers to utilise a common data fabric software platform that has connectivity to a wide range of data sources encompassing IoT devices at the edge, data streaming services, cloud storage, cloud and on-premises DBMSs and SaaS applications. Data fabric software should also be able to automatically infer schema from these sources and integrate with a data catalog so data engineers can search to see what source data is available to them to create data products and where sensitive data is so they can anonymise it as it is processed in a pipeline.

It also needs to enable pipelines to be designed that can automatically accommodate change and govern and protect data

It should also be possible for teams to collaboratively develop scalable, DataOps pipelines that can process structured, semi-structured and unstructured data with the ability to accommodate changes in schema, semantics, and infrastructure. Companies also need resilience to unexpected change and also the ability for pipelines to handle streaming data, change data capture and batch-oriented extract. Also, to accommodate complex transformations, and enable data governance policies to be enforced during data integration, people need a platform that is extensible, and that can automatically detect, mask, and encrypt sensitive data and integrate with other governance tools.

StreamSets DataOps Platform can do all of these things as well as support automated testing and deployment via CI/CD support. For this reason, it should be on any shortlist to help companies shorten time to value in a data driven enterprise.

About Intelligent Business Strategies

Intelligent Business Strategies is an independent research, education, and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, machine learning, advanced analytics, data management, big data, and enterprise business integration. Together, these technologies help an organisation become an *intelligent business*.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, data architecture, data governance, technology selection, and enterprise architecture. Mike is also conference chairman of Big Data LDN, the fastest growing data and analytics conference in Europe. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS, and European Managing Director of Database Associates. He teaches popular master classes in Data Warehouse Modernisation, Big Data Fundamentals, Centralised Data Governance, of a Distributed Data Landscape, Creating Data Products in a Data Lake, Lakehouse or Data Mesh for Use in Analytics, Machine Learning and Advanced Analytics, Real-time Analytics, and Data Virtualisation.

Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-Mail: info@intelligentbusiness.biz

The Value of Enterprise Data Fabric in a Data-Driven Enterprise