Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

The Technology of the Business Data Lake

Table of Contents

Overview 3

Business Data Lake Architecture 5

Designing the Business Data Lake 11

Conclusion 15

Overview

A new approach to providing data to all constituents of the enterprise, consolidating existing data marts to satisfy enterprise reporting and information management requirements

Many organizations have built enterprise data warehouses (EDWs) to meet their business's operational and reporting needs. Most EDW platforms are relatively expensive, costing upwards of $25,000 for 1TB of data storage, although costs have come down and computing power increased over the years.

Now, however, alternative technologies have matured to become viable cost-efficient alternatives to EDW processing. These technologies offer new ways to get value from enterprise data by providing users with the data views they really need, instead of a watered-down canonical view.

The Business Data Lake approach, enabled by Pivotal technology, reduces the complexity and processing burden on the EDW while preserving end-user interfaces and interactions with existing EDWs. Compared with a traditional EDW the approach delivers a significant cost advantage and improves the ability to respond to the needs of the business, while at the same time extending the life of EDW systems.

Introduction

The Pivotal Business Data Lake is a new approach to providing data to all constituents of the enterprise, consolidating existing data marts to satisfy enterprise reporting and information management requirements. Pivotal provides tools you can use both to create a new Business Data Lake and to extend the life of existing EDW solutions.

The Pivotal Business Data Lake also resolves a longstanding challenge in the area of operational reporting: the frequent conflict between the local reporting needs of individual business units and enterprise-wide reporting needs. With the Pivotal approach, there is no longer an issue of individual versus enterprise-wide views: individual business units can each get the local views they require, and there is also a global view to meet enterprise-wide needs. This is possible because Pivotal has combined the power of modern business intelligence (BI) and analytics into an integrated operational reporting platform that can be leveraged across the entire enterprise.

In addition, the Pivotal approach addresses concerns about the rising costs of EDWs versus the value they provide. The Pivotal Business Data Lake lowers costs by optimizing the data within an EDW, and provides more value by adding big data analytics into the EDW without the cost of scaling the EDW to process big data volumes[1].

Pivotal can help your organization to satisfy evolving information needs while handling new challenges such as big data processing and data access by mobile users and transactional systems. The Pivotal Business Data Lake adds performance to EDWs by providing lightning-fast, real-time, in-memory access for key information. This

1 Big data volumes: Any data over 1 petabyte in size or over 1 billion rows

means mobile users and transactional systems can leverage the power of your EDW without the cost of scaling traditional EDWs to meet transactional demands.

The Pivotal Business Data Lake supports line-of-business solutions with a single platform that also addresses enterprise needs. For operational reporting, it provides three key capabilities:

- The ability to rapidly ingest information from source systems
- The ability to create standard and ad hoc reports
- The ability to add real-time alert management

EDW pain points

Your organization is likely to encounter several challenges when trying to create or enhance EDWs to support requirements such as a single view of the customer:

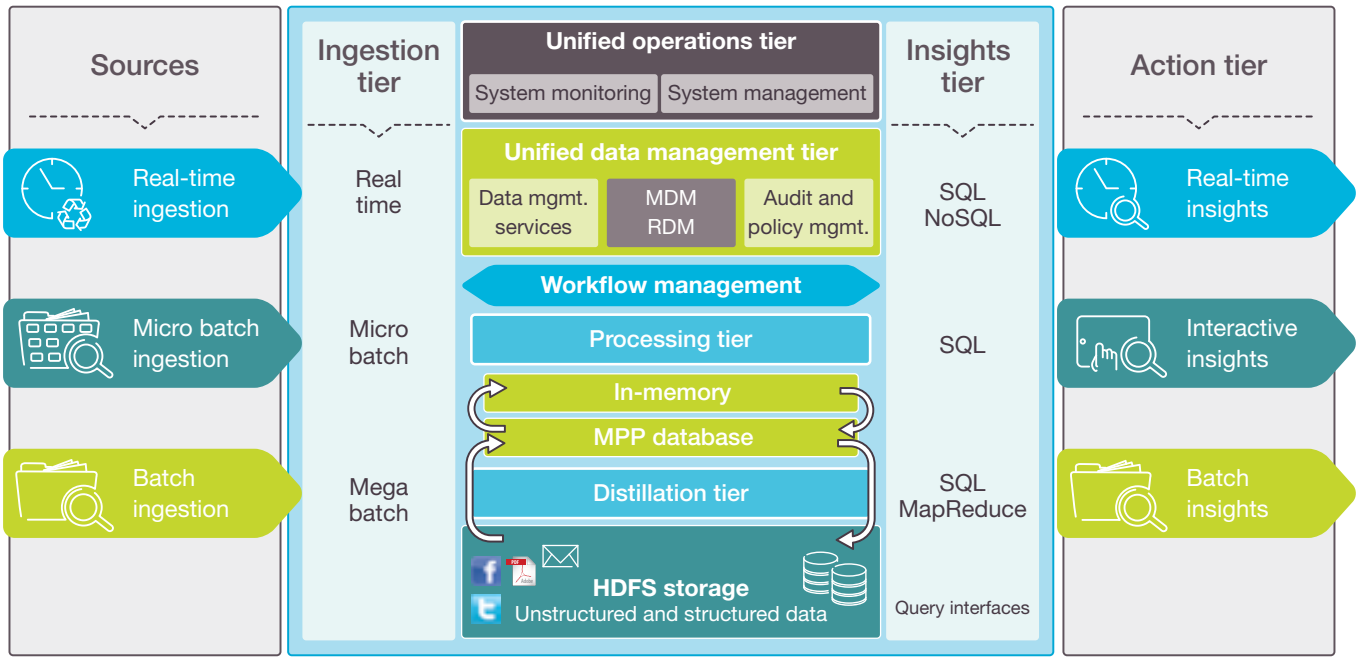
1. **Reconciling conflicting data needs.** Individual business units often need to enhance/extend global definitions to meet specific local needs. These enhancements/extensions may require additional data elements specific to the business unit, which may not be relevant from a corporate/global perspective. Because EDW implementations mandate the creation of a single consistent view of information across the enterprise, the conflicting views of the individual business units and the enterprise as a whole can be a problem. The Pivotal approach reconciles these conflicts by providing both global and local views of the same data.
2. **Providing real-time access.** EDWs are usually segregated from transactional and operational systems, which results in an inherent delay in information availability and data freshness. However, today's business decisioning systems need access to real-time information in addition to historical information to enable optimum decisions, better service, and product differentiation. Pivotal makes that possible through the use of performance-enhancing techniques like in-memory storage.
3. **Assembling data from multiple sources.** Data workers need an easy way to access the information required for their analysis. Usually the information is available in disparate systems. Each data worker has a preferred platform for analysis and sometimes the type of analysis dictates the environment. Today, data workers spend a lot of time getting the right data for analysis onto the appropriate analytic platform. With Pivotal's approach, it's easy to ingest data from multiple systems into a single analytics environment.
4. **Supporting ad hoc analysis.** In addition to regular operational reporting, enterprises need the ability to run ad hoc analysis from time to time. Operational systems typically are not capable of analysis without an adverse effect on performance. The parallelism of Pivotal Business Data Lake's architecture overcomes the constraints of the operational systems and makes it possible to run ad hoc analysis as needed.

To understand in more detail how Pivotal addresses these challenges, let's review the blueprint for building a Business Data Lake, and then see how applications can take advantage of it.

Business Data Lake Architecture

The figure below shows the key tiers of a Business Data Lake. Data flows from left to right. The tiers on the left depict the data sources, while the tiers on the right depict the integration points where insights from the system are consumed.

Figure 1: Business Data Lake Architecture



This figure also depicts the timeliness of data. The lower levels of the figure represent data that is mostly at rest, while the upper levels depict real-time transactional data that needs to flow through the system with as little latency as possible.

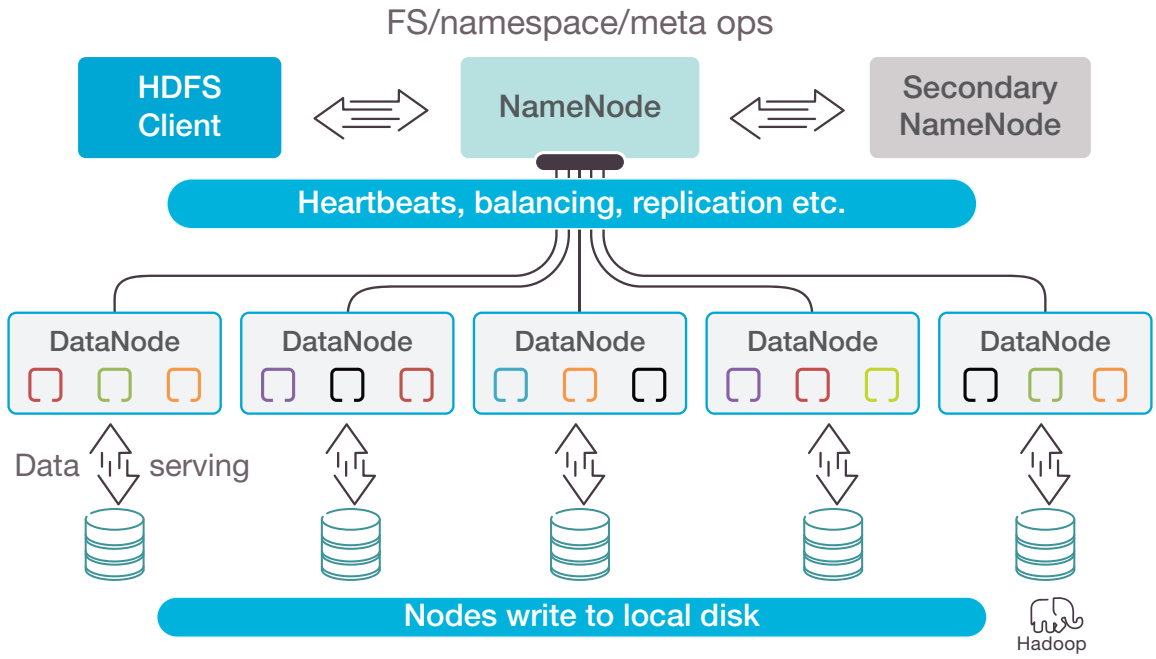
Unified operations that apply to all data – such as auditing and policy management, systems management, and the workflow that manages how data moves between the tiers – are represented separately in the figure.

Next, let's take a closer look at some of the key elements of the Business Data Lake.

Data storage tier

Different applications impose different requirements on the storage infrastructure. One class of application requires a real-time response to data access requests, while another class requires access to all historical data. A holistic approach to data needs storage of all the data plus real-time access to selected data.

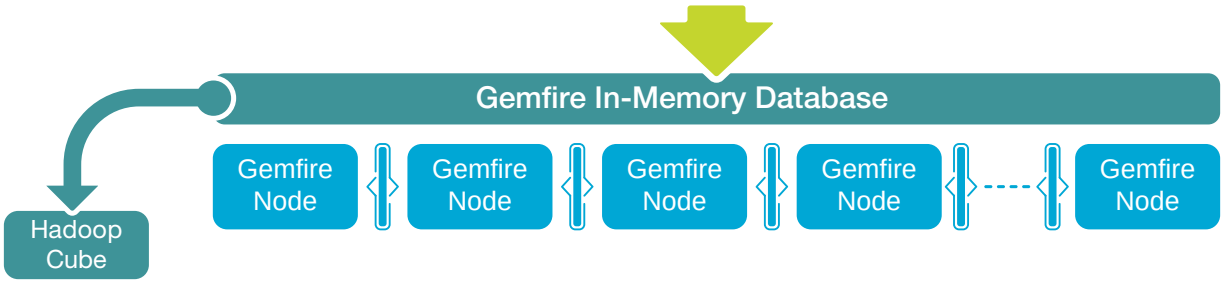
Figure 2: HDFS storage



The current explosion of both structured and unstructured data demands a cost-effective, reliable storage mechanism. The Hadoop Distributed File System (HDFS[2]) has emerged as the predominant solution, providing a low-cost landing zone for all data that is at rest in the system. One of the key principles of Hadoop is the ability to store data "as is" and distill it to add the necessary structure as needed. This "schema-on-read" principle eliminates the need for heavy extract transform load (ETL) processing of data as it is deposited into the system.

Support for real-time responses. Many systems need to react to data in real time. For these systems, the latency of writing the data to disk introduces too much delay. Examples include the new class of location-aware mobile applications, or applications that have to respond to events from machine sensors. For these systems, an in-memory data solution provides the ability to collect and respond to data with very low latency while it is in motion, and to persist the data on HDFS when at rest.

Figure 3: Support for real-time responses



Distillation tier

Many factors influence the location of data processing. On the one hand, data workers have their preferred access interfaces; on the other, the way data is stored also makes one access interface preferable over others. This interface gap needs to be bridged, which may require data movement and sometimes data transformation.

Here the Business Data Lake differs from traditional EDW solutions. With the Pivotal approach, the process of ingesting, distilling, processing, and acting upon the data does not rely on a pre-ordained canonical schema before you can store data. Instead, raw data can be ingested and stored in the system in its native form until it is needed. Schema and structure is added as the raw data is distilled and processed for action using MapReduce jobs.

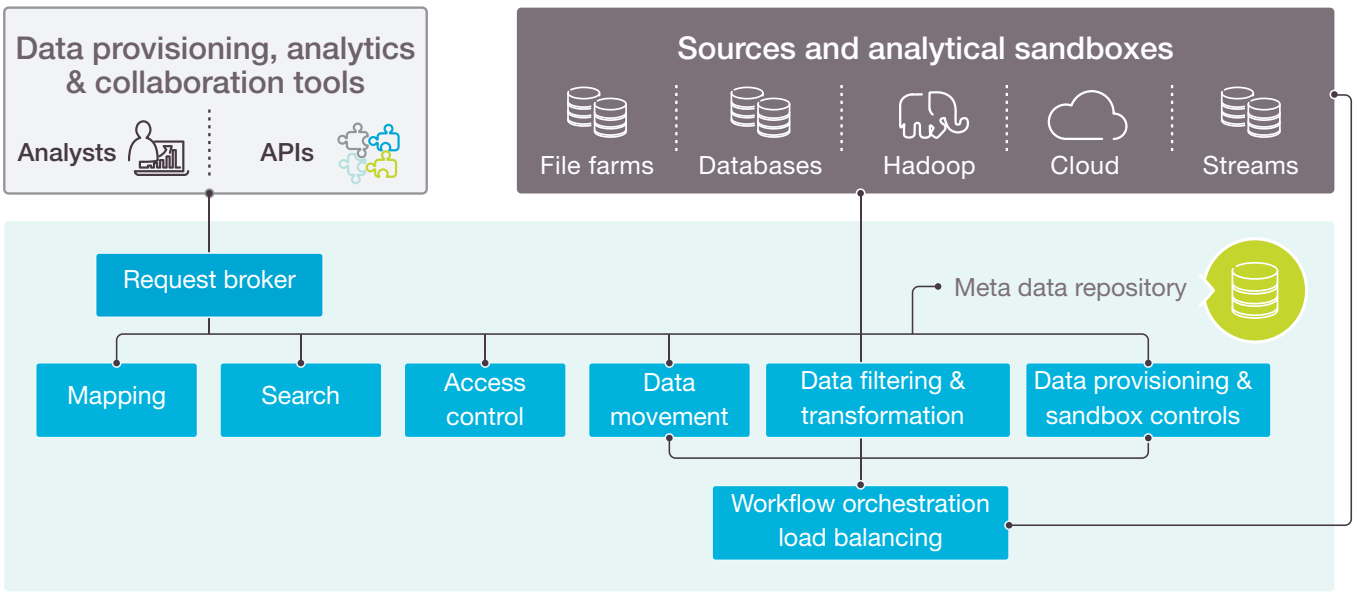
As raw data starts to take on more structure through the distillation process, analytical or machine learning algorithms can be applied to the data in place to extract additional insights from it. Individual events or transactions can also be streamed into the in-memory portion of the data lake, and, thanks to the system's grid nature, can be processed with very low latency.

The results of these real-time processes can be asynchronously moved to HDFS to be combined with other data in order to provide more insights for the business, or trigger other actions.

Unified data management tier

To manage and provide access to all the data that is collected in the Business Data Lake, authorized data workers can access data sets through a self-service portal that allows them to look through a metadata catalog of the data in the system and create a single view of data from across the company. Workflows in the system allow users to lease sandboxes of data and start complex analytics processes. When their policy-controlled leases expire, the resources are released back to the system until they are needed again.

Figure 4: Unified data management tier



2 http://yoyoclouds.wordpress.com/2011/12/

Insights tier

Insights from the Business Data Lake can be accessed through a variety of interfaces. In addition to Hadoop query interfaces like Hive or Pig, SQL – the lingua franca of the data world – can be used. Interactive analytics tools and graphical dashboards enable business users to join data across different data sets and draw insights.

To satisfy the needs of the data scientist community, MADlib analytic libraries can perform mathematical, statistical, and machine learning analysis on data in the system. Real-time insights can also generate external events – for example they can trigger actions in areas like fraud detection, or send alerts to mobile applications.

Components for the Business Data Lake

Having reviewed the Business Data Lake architecture, we'll now consider the enterprise-class products from Pivotal that you can integrate to create your own Business Data Lake. The table below summarizes these products.

Product	Description
Greenplum Database	A massively parallel platform for large-scale data analytics warehouses to manage and analyze petabytes of data – now available with the storage tier integrated on Hadoop HDFS (with HAWQ query engine). It Includes MADlib, an in-database implementation of parallel common analytics functions.
GemFire	A real-time distributed data store with linear scalability and continuous uptime capabilities – now available with storage tier integrated on Hadoop HDFS (GemFire XD).
Pivotal HD	Commercially supported Apache Hadoop. HAWQ brings enterprise-class SQL capabilities, and GemFire XD brings real-time data access to Hadoop.
Spring XD	Spring XD simplifies the process of creating real-world big data solutions. It simplifies high-throughput data ingestion and export, and provides the ability to create cross-platform workflows
Pivotal Data Dispatch	On-demand big data access across and beyond the enterprise. PDD provides data workers with security-controlled, self-service access to data. IT manages data modeling, access, compliance, and data lifecycle policies for all data provided through PDD.
Pivotal Analytics	Provides the business community with visualizations and insights from big data. Data from different sources can be joined to create visualizations and dashboards quickly. Pivotal Analytics can infer schemas from data sources, and automatically creates insights as it ingests data, freeing up business analysts to focus on analyzing data and generating insights rather than manipulating data.

The next table relates these Pivotal products to specific Business Data Lake tiers. Often, there are multiple products to support a given tier, sometimes with overlapping capabilities, so you need to pick the appropriate product for your business requirements. The discussion following the table will help with these choices.

Tiers	Description
Storage	*Ability to store **all** (structured and unstructured) data cost efficiently in the Business Data Lake* Pivotal HD: HDFS is the storage protocol on which the industry is standardizing for all types of data.
Ingestion	*Ability to bring data from multiple sources across all timelines with varying Quality of Service (QoS)* GemFire XD: Ideal platform when **real-time** performance, throughput and scalability are crucial. Spring XD: Ideal platform when throughput and scalability are critical, with **very good latency**. Pivotal HD: Flume and Sqoop are some of the open source ingestion products. Ideal when throughput and scalability are critical with **reasonable** latency expectations.
Distillation	*Ability to take data from the storage tier and convert it to structured data for easier analysis by downstream applications* Pivotal Data Dispatch: Ideal **self-serve** platform to convert data from the ingested input format to the analytics format. Essentially, it is about running ETL to get the desired input format. Pivotal Analytics: Provides analytics insights such as text indexing and aggregations on data ingest. ETL products: Clients can also use industry-standard ETL products such as Informatica or Talend to transform data from the ingested input format to the analytics format.
Processing	*Ability to run analytical algorithms and user queries with varying QoS (real-time, interactive, batch) to generate structured data for easier analysis by downstream applications* Pivotal HD: Ability to analyze **any type of data** using the Hadoop interfaces for data analysis such as Hive, HBase, Pig and MapReduce. HAWQ: Process complex queries and respond to user requests in **interactive time**. GemFire XD: Process queries and respond to user request in **real time**. Spring XD: Managing **cross-platform workflows**.

Tiers	Description
Insights	*Ability to analyze all the data with varying QoS (real-time, interactive, batch) to generate insights for business decisioning* Pivotal HD: Extract insights from **any type of data** using the Hadoop interfaces for data analysis, such as Mahout, Hive, HBase, Pig and MapReduce. HAWQ: Extract insights in **interactive time** using **complex analytical** algorithms. GemFire XD: Extract insights in **real time** from data stored in memory.
Action	*Ability to integrate insights with business decisioning systems to build data-driven applications.* AppFabric: Redis and RabbitMQ are used to integrate with existing business applications. New business applications can use Spring or other products. GemFire: Continuous query mechanism provides CEP-like capability to react to events as they happen. Pivotal CF: Improves application development velocity by simplifying the deployment of applications to your public or private cloud.
Unified data management	*Ability to manage the data lifecycle, access policy definition, and master data management and reference data management services* Pivotal Data Dispatch: Enables IT to define metadata centrally for data workers to find and copy data in the sandbox environment. However, master data management and reference data management services are capabilities not currently available from the Pivotal data fabric products.
Unified operations	*Ability to monitor, configure, and manage the whole data lake from a single operations environment* Pivotal Command Center: Unified interface to manage and monitor Pivotal HD, HAWQ and GemFire XD[3]. Pivotal is continuing to improve unified operations across the platform.

Designing the Business Data Lake

There are many architectural tradeoffs to consider in designing a Business Data Lake. Pivotal's products work together to help you build a solution that meets your specific needs. In this section we'll dive deeper into the architecture and explain how individual components fit in, with some comments as to which options may work best for a specific business requirement.

Data ingestion using Pivotal products

One of the principal differences between a data lake and a traditional EDW approach is the way data is ingested. Rather than performing heavyweight transformations on data to make it conform to a canonical data model, the data can be ingested into the data lake in its native form. It is important to think of data ingestion in terms of batch size and frequency – data arriving into a system can be grouped into mega batches, micro batches, and streams of real-time data. The Pivotal product portfolio has ways to deal with each of these groups.

Ingesting real-time data ("streaming"). Real-time data needs to be collected from devices or applications as it is generated, one event at a time. Much critical streaming enterprise data is revenue bearing – for example, credit card transactions – so data quality, reliability and performance are vital. They can be assured by ingesting transactions with Pivotal GemFire, which supports traditional XA transactions, and keeps multiple redundant copies of data to provide fault tolerance, while also achieving great performance. There is also a growing segment of streaming data where extreme scalability is more important than data quality and reliability: for example, sensor data or page view data. For this data, GemFire supports configurable levels of reliability, allowing you to maximize performance, or balance performance and reliability, to meet your needs. Where extreme scalability isn't a requirement, Spring XD can be an excellent choice for streaming real-time data. It uses in-memory technology for great performance, but also focuses on making it easy to configure inputs and outputs to implement enterprise integration patterns. A Pivotal representative can help you determine whether GemFire or Spring XD is better for your specific application.

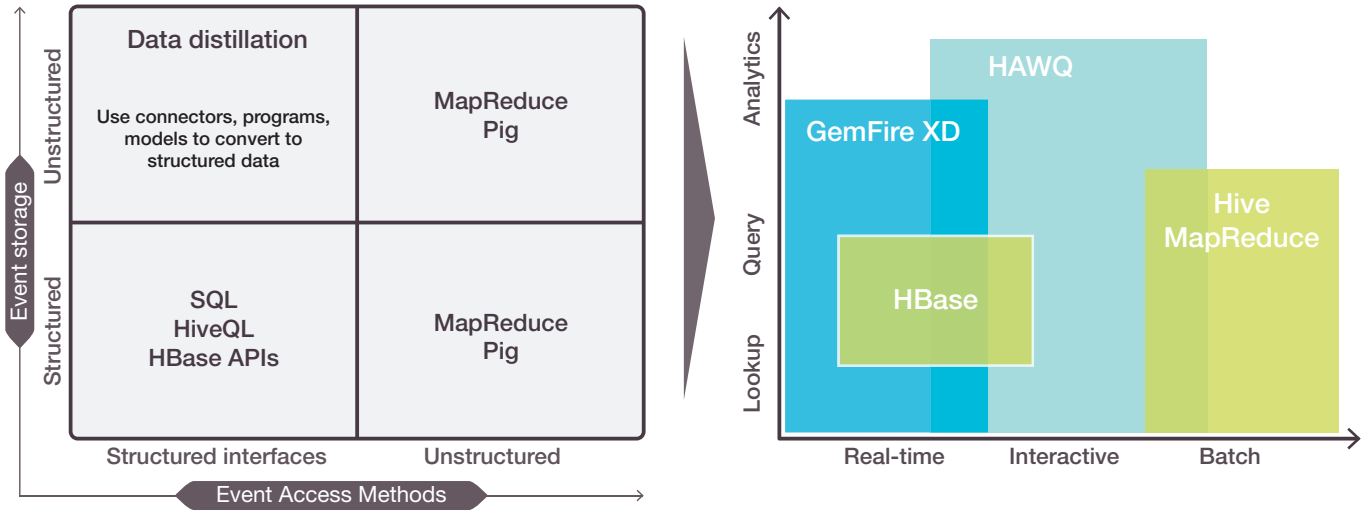
Ingesting batches of data. Spring XD excels in use cases where small chunks of data are batched up and pushed between systems at a regular frequency. It provides transaction management, chunking of data, and a restart mechanism built specifically for batch processing. Custom processing can be executed on batches while the data is still being ingested, before it is stored on HDFS. The processing can be as simple as transformations and lookup, or as complex as machine learning, scoring or address cleanup.

Bulk data ingest. Often, large amounts of data need to be moved from one platform to another. A key to success in bulk data transfer is maximizing network bandwidth without impacting other applications that share the same network resources. Pivotal DataLoader is designed specifically for this: it can ingest data at wire speed, but also includes a throttling capability to ensure other applications aren't affected. DataLoader also provides the ability to restart or resume the transfer process if it detects a failure.

3 GemFire Integration with Command Center is on the product delivery roadmap

The two diagrams below depict the spectrum of data loading challenges, and the tools in the Pivotal Business Data Lake portfolio that address them.

Figure 5: Spectrum of data loading challenges



Data distillation and processing with Pivotal products

Data distillation and processing are related but separate topics in a data lake environment. Distillation is about refining data and adding structure or value to it. Processing is about triggering events and executing business logic within the data tier.

Distillation. When accessing unstructured data in the Business Data Lake, standard Hadoop MapReduce jobs can distill it into a more usable form in Pivotal HD. The same MapReduce jobs can also access HAWQ and GemFire XD data. Examples of distilling data to add structure include complex image processing, video analytics, and graph and text analytical algorithms. All of them use the MapReduce interfaces to generate insights from unstructured data. Pivotal HD also supports standard Hadoop ecosystem tools like Pig and Hive, providing a simpler way to create MapReduce jobs that add structure to unstructured data.

Processing. It is important that real-time events can be acted upon in real time. For example, a mobile app that offers a free cup of coffee to people who walk by a store must do so instantly if it is to be effective. This real-time action can be achieved using Pivotal GemFire, which stores data in the memory of a grid of computers. That grid also hosts event handlers and functions, allowing client applications to subscribe to notifications for "continuous queries" that execute in the grid. With GemFire, you can build the type of applications normally associated with complex event processing packages.

Pivotal products for insights

To gain insight from your Business Data Lake, you need to provide access through a standard SQL interface: this way, you can use existing query tools on big data. Pivotal's HAWQ component is a full SQL query engine for big data, with a distributed cost-based query optimizer that has been refined in the Greenplum database to maximize performance for big data queries.

Insights with significant business value usually need to be delivered in real time. Responding to news in high-frequency trading systems, finding fraudulent activity in financial areas, identifying intruders from the access log, performing in-session targeting – these are just a few examples of real-time insights. GemFire's complex event processing and data alerting ability generates these insights, which can then be acted on by automated applications. Spring XD also has capabilities for running simple analytics to get insights while data is in motion.

Pivotal Analytics provides the business community with insights from big data, joining data from different sources to create visualizations and dashboards quickly. A key feature is it can infer schemas from data sources, automatically creating insights like time series analysis and text indexes so that business analysts can spend most of their time working with insights rather than manipulating data.

For even more sophisticated data analytics, Pivotal supports the MADlib library directly inside the data tier, providing scalable in-database analytics. There are data-parallel implementations of mathematical, statistical and machine-learning methods for structured and unstructured data. Pivotal's data tier also provides the ability to run R models directly inside the database engine, so your models can be parallelized and can access data in place.

Taking action on data with Pivotal products

A Business Data Lake implementation enables enterprise users to ingest data, manage data, and generate insights from data. Putting these insights into action requires new applications, however. The Spring Tool Suite helps you build new big-data driven applications rapidly; you can then integrate them with your business decisioning systems. In addition, integration components such as Spring XD and RabbitMQ enable you to integrate the insights from your data into existing business applications across the enterprise. Pivotal CF provides a new-generation application container for the cloud that increases development velocity by dramatically decreasing the time it takes to deploy applications to your public or private cloud.

Managing data with Pivotal products

A Business Data Lake enables you to keep all data on a single storage platform, and to achieve flexibility while maintaining a stable global perspective. All of this requires the ability to find and share data among business users. With Pivotal Data Dispatch, your IT team can make selected data available for sharing in accordance with your access policies. Business users can also find a data set they're interested in and bring it on demand to the platform that provides their preferred interfaces for data analysis. Business users and IT teams can specify the transformations and processing required on the data as it is moved.

Unified operations in the Pivotal product Stack

The Business Data Lake requires unified monitoring and manageability for data, users, and the environment. The Pivotal HD and HAWQ interfaces are now integrated for operability and manageability, while GemFire XD, Pivotal Data Dispatch and DataLoader currently have separate interfaces. As Pivotal's manageability unification strategy progresses, components of the Business Data Lake will be managed from the unified interfaces.

Maximizing flexibility

For the Business Data Lake approach to succeed, it's crucial to offer flexibility at local level while providing a certified global view of the data. You can use various design patterns, some of which are outlined below, to provide a standard schema for the global view and enable flexibility for extensions/enhancements in the local view of the same data.

Real-time global integration

GemFire XD enables a design pattern where global information is kept in a database that provides a real-time view. Local information is stored locally, without impacting the global certified data.

Dealing with variable data quality

It is essential to understand the quality of the data that your business bases its decisions on. Data quality issues show up in various ways:

- Incomplete data – some data either is missing or arrives late
- Invalid data – there is bad or otherwise erroneous data
- Uncertainty – you don't know how much of the data is accurate

These pointers can give directional guidance to business users; however, this may not be enough in the case of financial applications, where stringent controls and validations are required.

Keeping most of the data available in real time, as is possible with GemFire XD, facilitates a consistent view of important data across the enterprise. Additionally, the quality of data can be indicated, so that users can take it into account before leveraging that data for business decisioning. Where appropriate, business processes can be triggered only when the data quality is 100% i.e. certified and approved.

Multiple views

HAWQ and GemFire XD are both capable of joining data at scale, a capability that can be leveraged when designing applications. An enterprise's users need to agree on the global fields, and strictly enforce standardization by placing all the global data in a base table. Further tables can then be derived from this one to provide local flexibility; these tables will store additional local information and relate it back to the base table. Local data workers then can join the tables to get the view they require for data analysis. Depending on the usage of the local extensions, the views can be materialized to optimize performance, or joined opportunistically as required.

Conclusion

The Pivotal Business Data Lake provides a flexible blueprint to meet your business's future information and analytics needs while avoiding the pitfalls of typical EDW implementations. Pivotal's products will help you overcome challenges like reconciling corporate and local needs, providing real-time access to all types of data, integrating data from multiple sources and in multiple formats, and supporting ad hoc analysis.

It combines the power of modern BI and analytics in an integrated operational reporting platform that will provide your entire enterprise with valuable insights to inform decision-making. By using this approach to make the most of your data, you can improve the performance of both new and existing EDW systems, and also extend their life.

Pivotal and Capgemini are co-innovating to bring market leading technology, best practices and implementation capabilities to our enterprise customers.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Pivotal™

Find out more at **www.capgemini.com/bdl**
and **www.gopivotal.com/businessdatalake**

Or
contact us at **bim@capgemini.com**

About Capgemini

With more than 130,000 people in 44 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2012 global revenues of EUR 10.3 billion.

Together with its clients, Capgemini creates and delivers business and technology solutions that fit their needs and drive the results they want. A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

MCOS_GI_MK_20131119