CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

Apex Institute of Technology

Department of Computer Science & Engineering

Introduction To Data Science (CST-291)

Dr. Jitender Kaushal
Associate Professor
CSE(AIT), CU

DISCOVER . LEARN . EMPOWER

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

COURSE OUTCOMES

On completion of this course, the students shall be able to:-

CO1	Identify and describe the importance of Big data analysis over Conventional Database management System.

Contents to be Covered

- Big Data Architecture
- Types of Big Data Architecture

Big Data Architecture

- A big data architecture is designed to handle the **ingestion, processing, and analysis of data** that is too large or complex for traditional database systems.

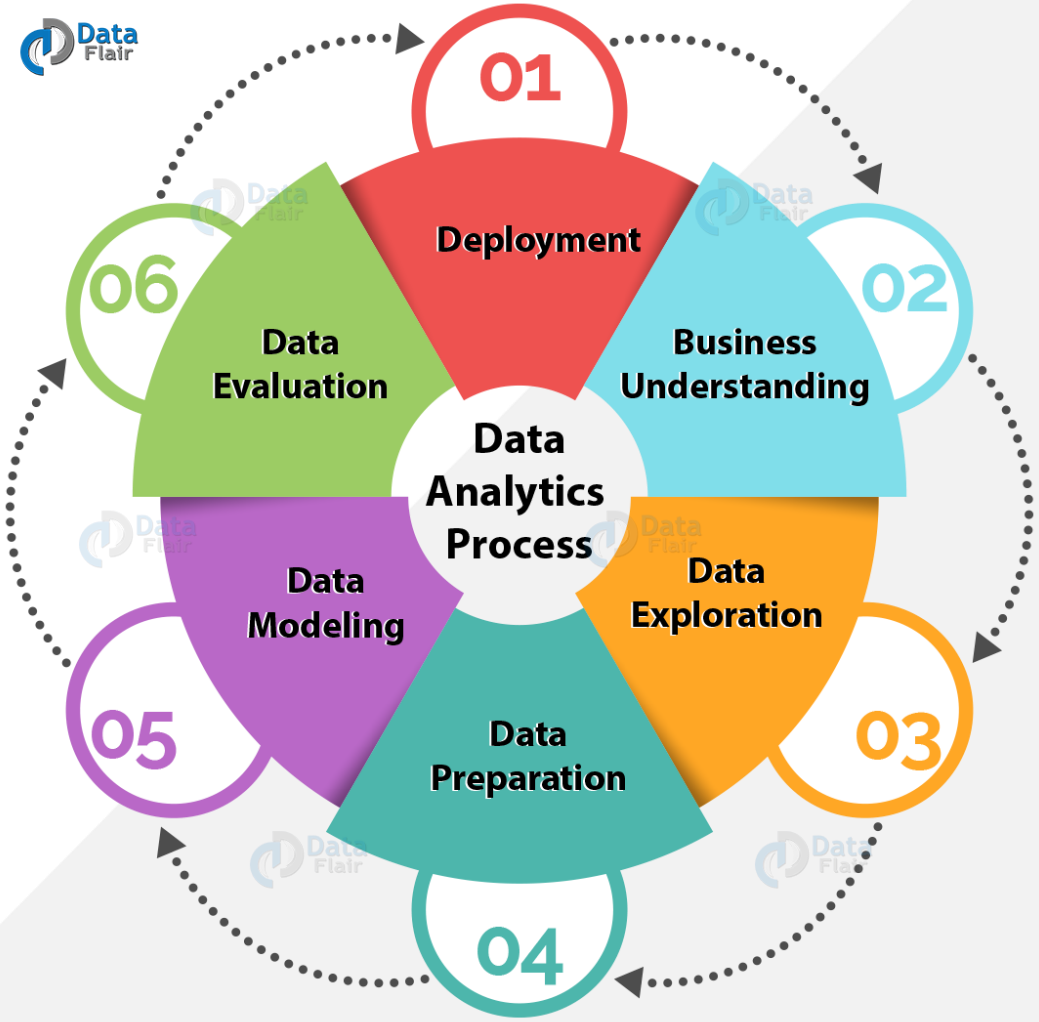
- Big data architecture is the **cardinal system** supporting big data analytics. The bedrock of big data analytics, **big data architecture is the layout that allows data to be optimally ingested, processed, and analyzed.**

- In other words, big data architecture is the **key player** that **drives data analytics** and provides a means by which **big data analytics tools can extract vital information** from obscure or doubtful data and drive **meaningful and strategic business decisions**.

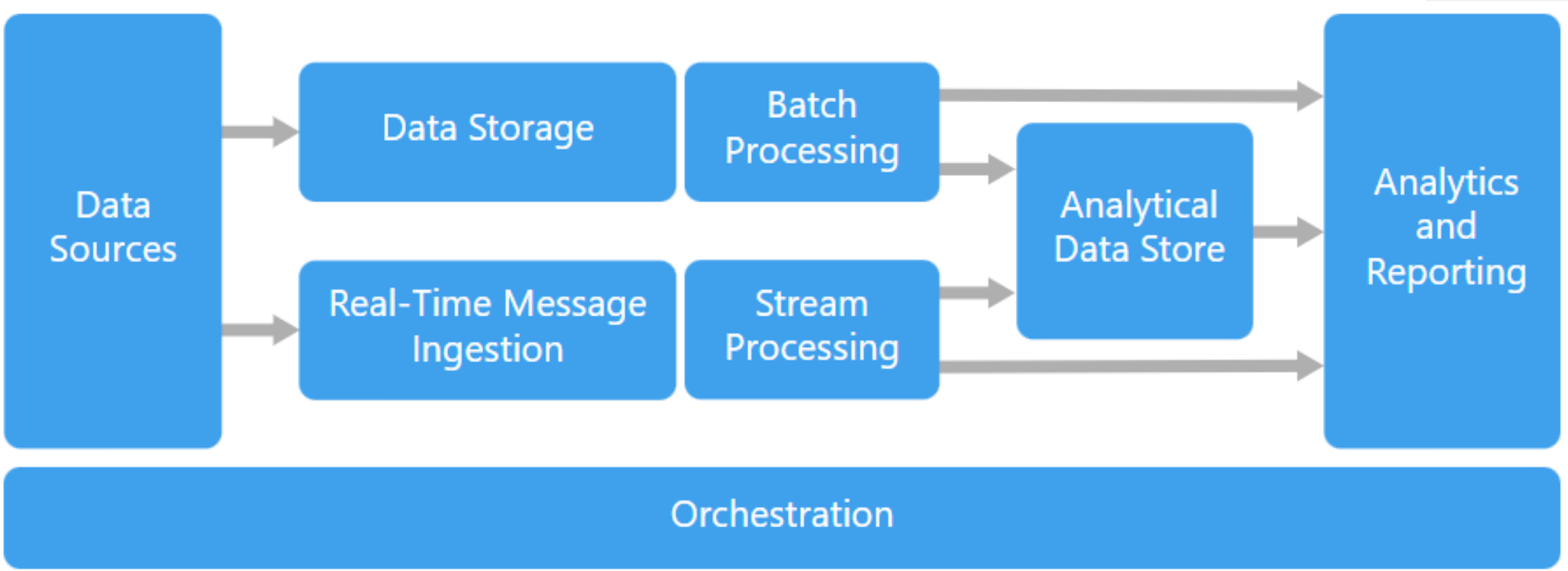
Big data ingestion **gathers data and brings it into a data processing system where it can be stored, analyzed, and accessed**.

Big data Architecture

Big data analytics **describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions**. These processes use familiar statistical analysis techniques—like **clustering and regression**—and apply them to more extensive datasets with the help of newer tools.



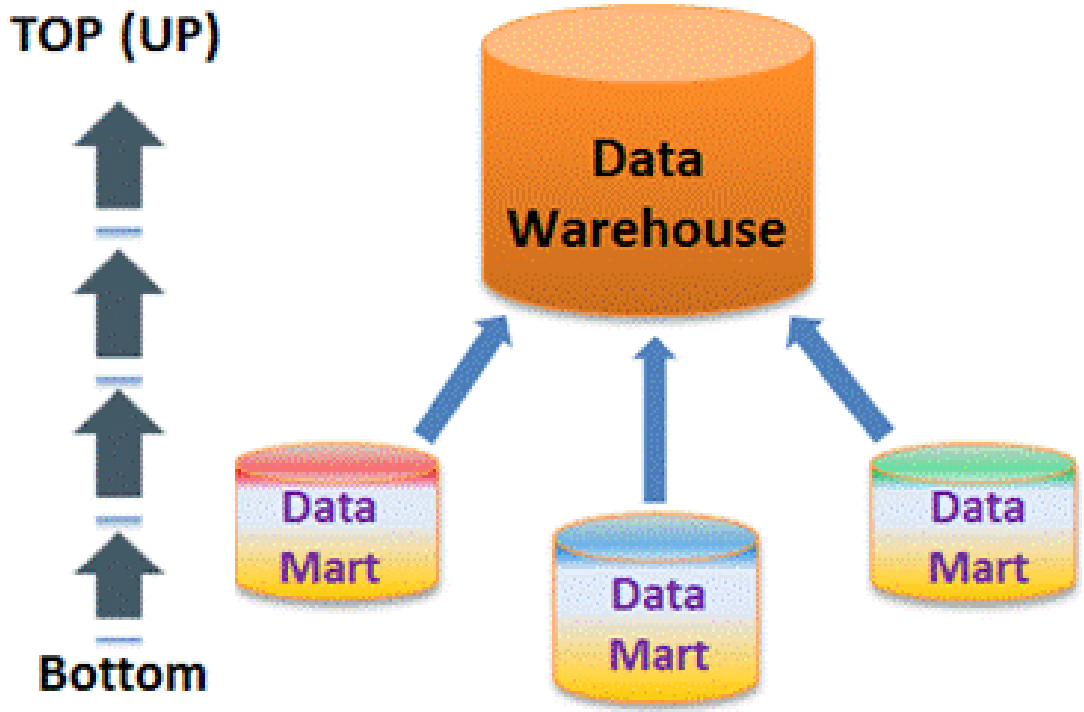
Big data Architecture



Components of big data architecture

- **Data sources:** The obvious **starting point** of all big data solutions **data sources** may be **static files** produced by applications **(web server log files)**, application data sources **(relational databases)**, or **real-time data sources (IoT devices)**.
- **Data storage:** Often referred to as a **data lake**, a **distributed file store** holds bulks of **large files** in different formats, which are subsequently used for **batch-processing operations.**
- **Batch processing:** In order to make **large datasets analysis-ready**, batch processing carries out the **filtering, aggregation, and preparation of the data files** through long-running batch jobs.
- **Message ingestion:** This component of the big data architecture includes a way **to capture and store messages from real-time sources** for **stream processing**.

- **Stream processing:** Another preparatory step before data analytics, stream processing **filters and aggregates the data** after capturing real-time messages.
- **Analytical data store:** After preparing the data for analytics, most big data solutions serve the processed data in a **structured format for further querying using analytical tools**. The analytical data store that serves these queries can either be a **Kimball-style relational data warehouse** (follows a bottom-up approach to data warehouse architecture design in which data marts (a subset of a directorial information store) are first formed based on the business requirements) or a **low-latency NoSQL technology**.



Kimball's Data Model Approach

- **Analysis and reporting: One of the critical goals** of most big data solutions, data analysis and reporting **provides insights into the data**. For this purpose, the big data architecture may have **a data modeling layer, support self-service**, or even incorporate **interactive data exploration**.
- **Orchestration:** An orchestration technology can **automate the workflows involved in repeated data processing operations**, such as **transforming the data source**, **moving data between sources and sinks**, **loading** the processed data into an **analytical data store**, and **final reporting**.

Orchestration is **the coordination and management of multiple computer systems, applications, and/or services, stringing together multiple tasks in order to execute a larger workflow or process**. These processes can consist of **multiple tasks** that are **automated** and can involve **multiple systems**.

Big data architecture layers

- The components of big data analytics architecture primarily consist of **four logical layers** performing key processes. The layers are merely logical and provide a means to organize the components of the architecture.
- **Big data sources layer:** The data available for analysis will **vary in origin and format**; the format may be **structured, unstructured, or semi-structured**, the **speed** of data arrival and **delivery** will vary according to the source, the **data collection** mode may be direct or through data providers, in **batch** mode or in real-time, and the location of the data source may be **external** or **within the organization**.
- **Data massaging and storage layer:** This layer acquires data from the data sources, **converts and stores** it in a format that is **compatible with data analytics tools** (e.g CSV, JSON files). **Governance policies and compliance regulations** primarily decide the suitable storage format for different types of data.

- **Analysis layer:** It extracts the data from the data massaging and storage layer (or directly from the data source) to **derive insights from the data**.

- **Consumption layer:** This layer **receives the output** provided by the analysis layer and presents them **to the relevant output layer**. The consumers of the output may be **business processes, humans, visualization applications, or services**.

Big data architecture process

- **Data source connection:** Fast and efficient data ingression demands seamless connectivity to different storage systems, protocols, and networks, achieved by **connectors and adapters**.
- **Big data governance:** Data governance operates right from data ingestion and continues through data processing, analysis, storage, archive or deletion, and includes provisions for **security and privacy**.
- **Management of systems:** Modern big data architecture comprises **highly scalable and large-scale distributed clusters**; these systems must be closely monitored through **central management consoles**.
- **Quality of service (QoS):** QoS is a **framework** that offers **support for defining the data quality, frequencies and sizes of ingestion, compliance policies**, as well as **data filtering**.

Big data architecture best practices

- Big data architecture best practices refer to **a set of principles of modern data architecture** that help in **developing a service-oriented approach** while at the same time addressing business needs in a fast-paced data-driven world.
- Align the **big data project** with the **business vision.**
- The big data project should be in line with the **business goals** and the **organizational context** with a clear understanding of the data architecture work **requirements, frameworks and principles** to be used, the key drivers of the organization, business technology elements currently in use, **business strategies and organizational models, governance and legal frameworks, and pre-existing and current architecture frameworks**.

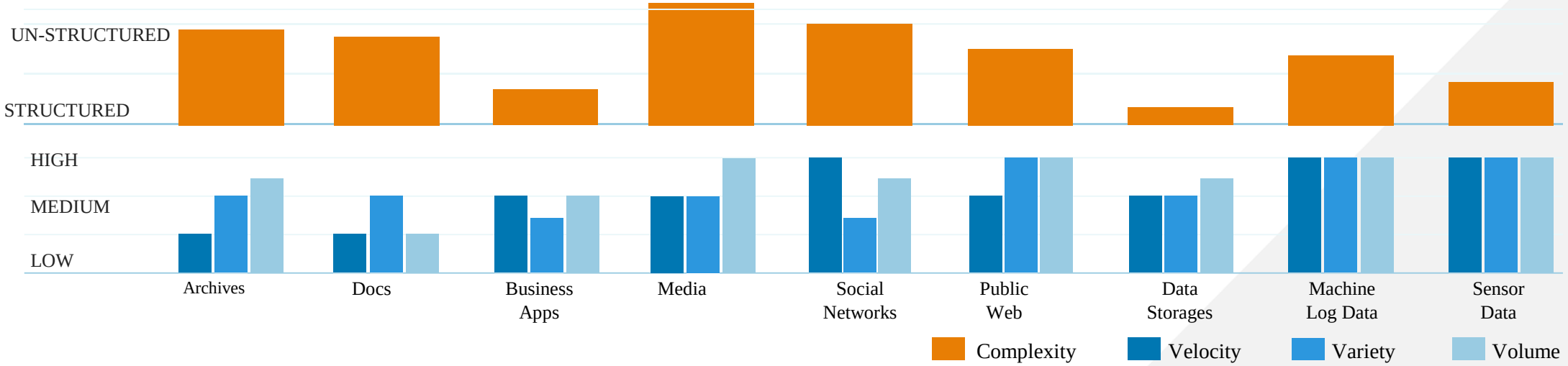
- **Identify and categorise data sources**
- For data to be normalised into a standard format, data sources must be **identified and categorised**. The categorisation may either be structured data or unstructured data; while the former is usually formatted through predefined database techniques, the latter does not follow a consistent and well-defined format.
- **Consolidate data into a single Master Data Management system**
- **Batch processing and stream processing** are two methods via which data can be consolidated for querying on demand. In this regard, it is imperative to mention that Hadoop is a popular, open-source batch processing framework for storing, processing, and analysing vast volumes of data. The Hadoop architecture in big data analytics consists of four components – MapReduce, HDFS (HDFS architecture in big data analytics follows the master-slave model for reliable and scalable data storage), YARN, and Hadoop Common. In addition, for querying, a relational DBMS or NoSQL database can be used for storing the Master Data Management System.

- **Provide a user interface that eases data consumption**
- An intuitive and customizable user interface of the big data application architecture will make it easier for the users to consume data. **For example**, it could be an SQL interface for data analysts, an OLAP (Online Analytical Processing Server) interface for business intelligence, the R-language for data scientists, or a real-time API for targeting systems.

(R is an advanced language that performs various complex statistical computations and calculations. Moreover, R interprets the data in a graphical form, making it easy to interpret and understand.)

- **Ensure security and control**
- Instead of enforcing data policies and access controls on downstream data stores and applications, it is done directly on the raw data. This unified approach to data security has been further necessitated by the growth of platforms such as Hadoop, Google BigQuery, Amazon Redshift, and Snowflake and made into a reality by data security projects like Apache Sentry.

Big Data Challenges



Archives

Scanned documents, statements, medical records, e-mails etc..

Docs

XLS, PDF, CSV, HTML, JSON etc.

Business Apps

CRM, ERP systems, HR, project management etc.

Media

Images, video, audio etc.

Social Networks

Twitter, Facebook, Google+, LinkedIn etc.

Public Web

Wikipedia, news, weather, public finance etc

Data Storages

RDBMS, NoSQL, Hadoop, file systems etc.

Machine Log Data

Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data

Smart electric meters, medical devices, car sensors, road cameras etc.

Big Data Analytics

	Traditional Analytics	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• Predictive analytics • Data Science
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	Causation: what happened, and why?		Correlation: new insight More accurate answers

Big Data Analytics Use Cases

Low Latency
Reliability

Consumers

Intelligent Agents

Volume
Performance

Data Scientists/
Analysts

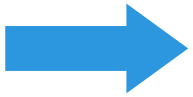
Data Quality
Self Service

Business Users

Big Data Analytics Reference Architectures

Architecture Drivers:

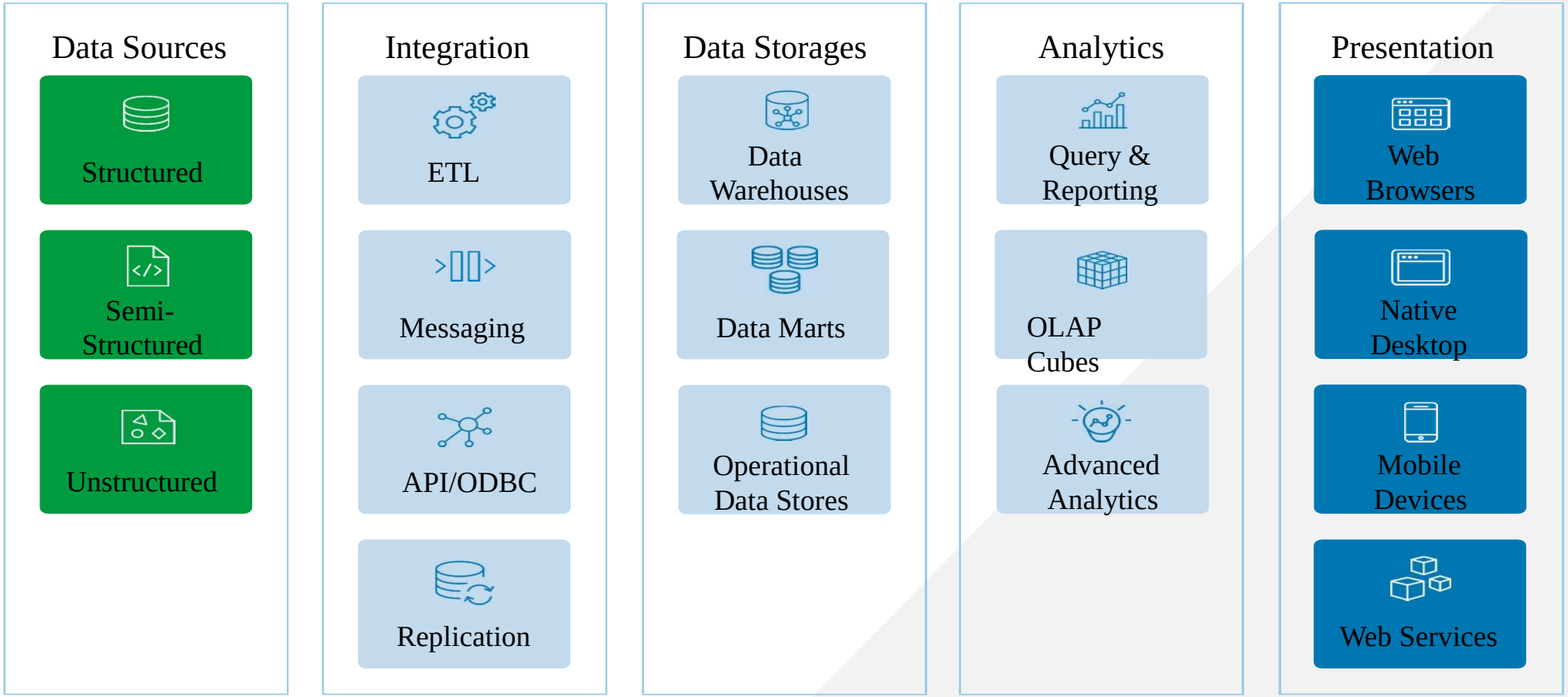
- Volume
- Sources
- Throughput
- Latency
- Extensibility
- Data Quality
- Reliability
- Security
- Self-Service
- Cost



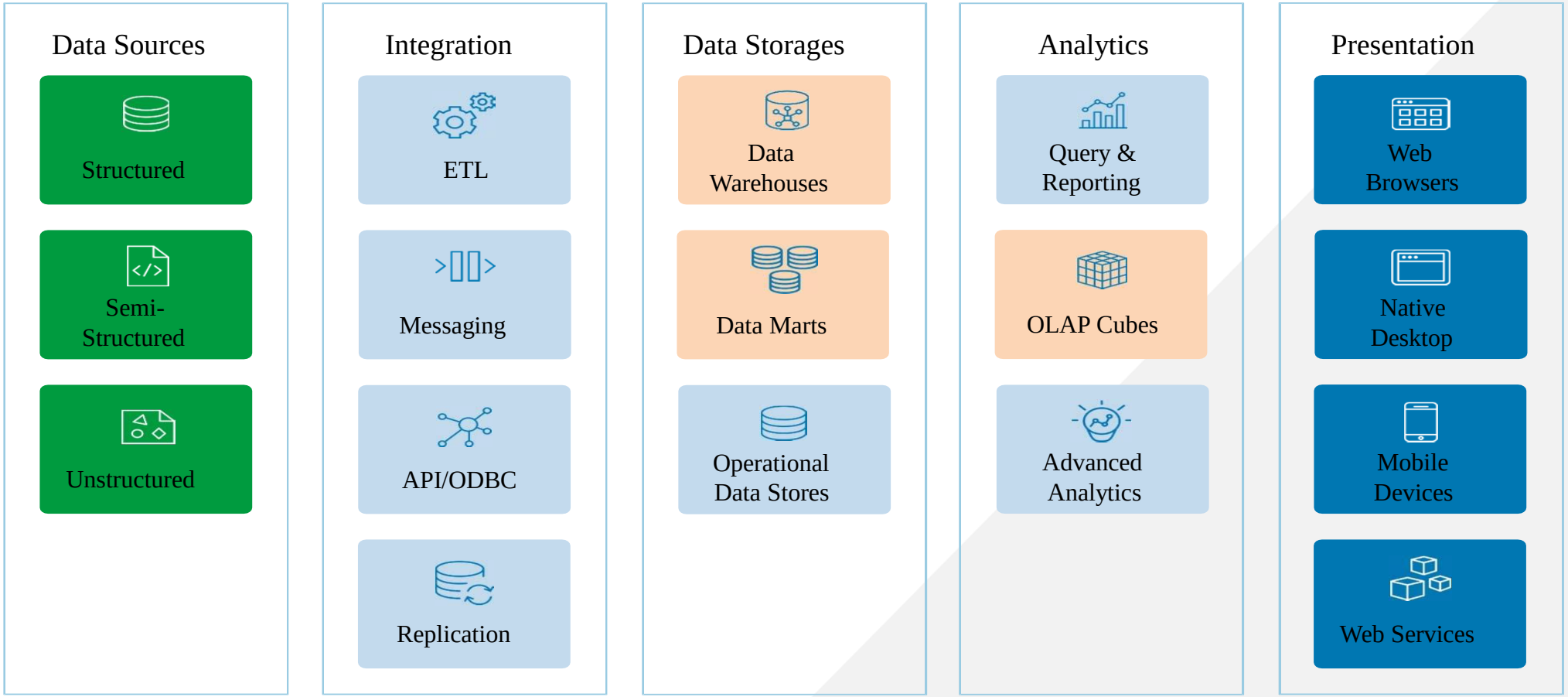
Reference Architectures:

- Extended Relational
- Non-Relational
- Hybrid

Relational Reference Architecture

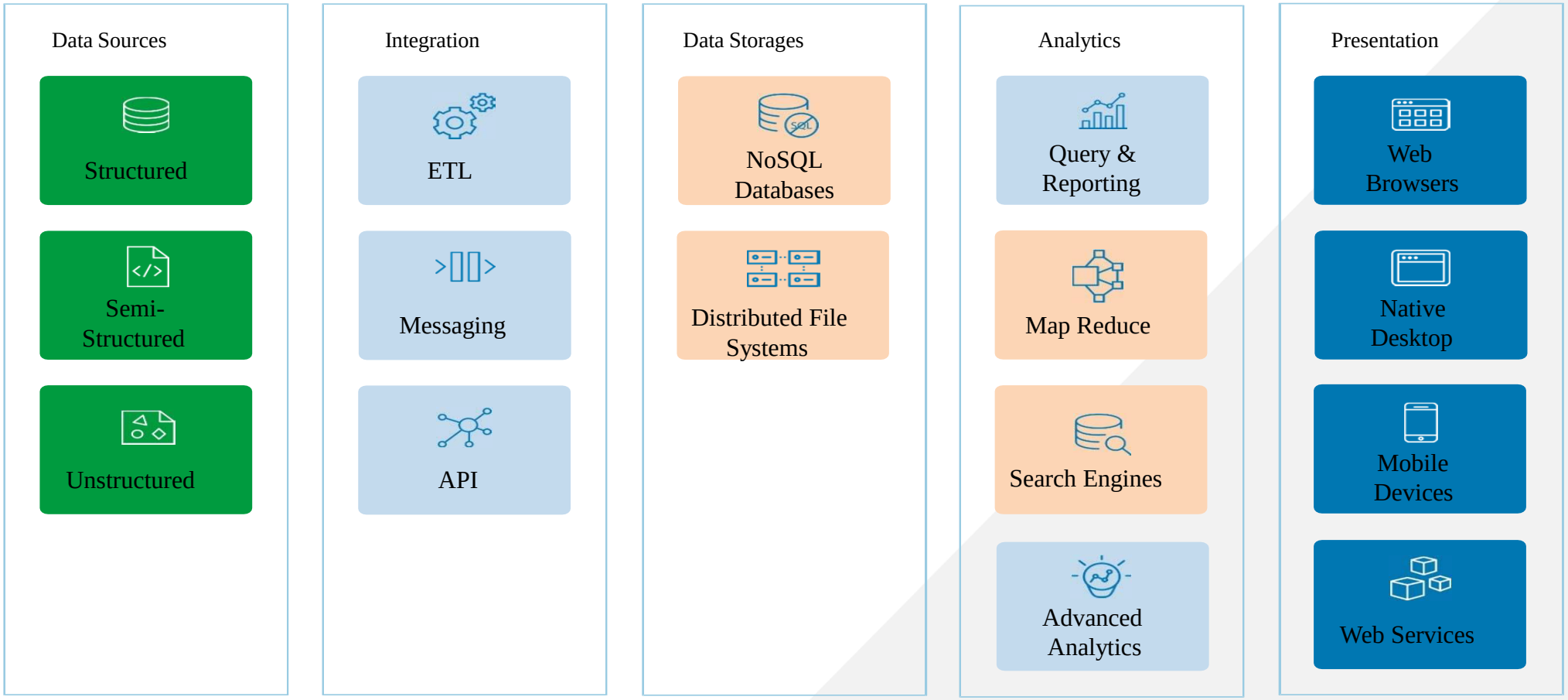


Extended Relational Reference Architecture



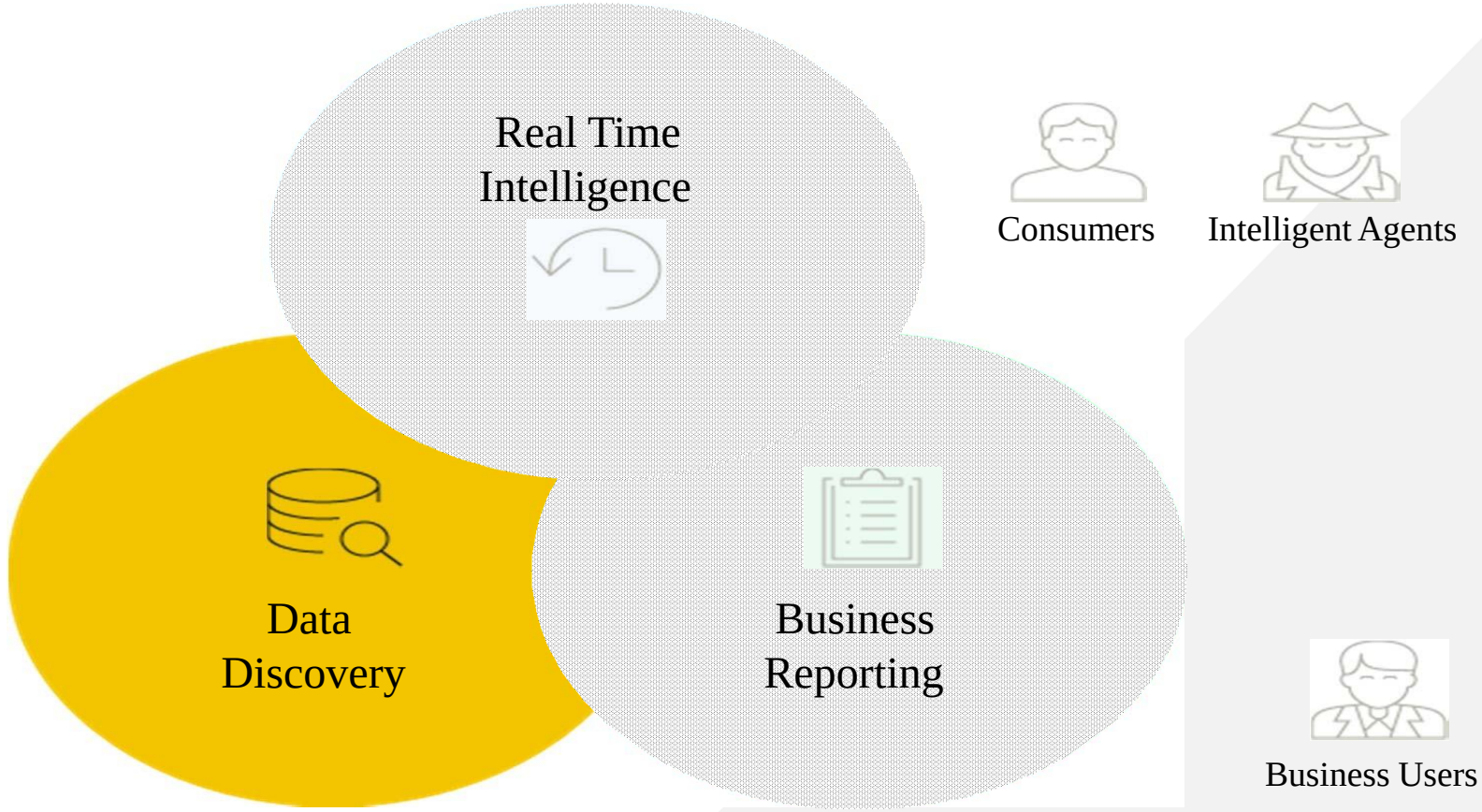
Key components affected with Big Data challenges

Non-Relational Reference Architecture

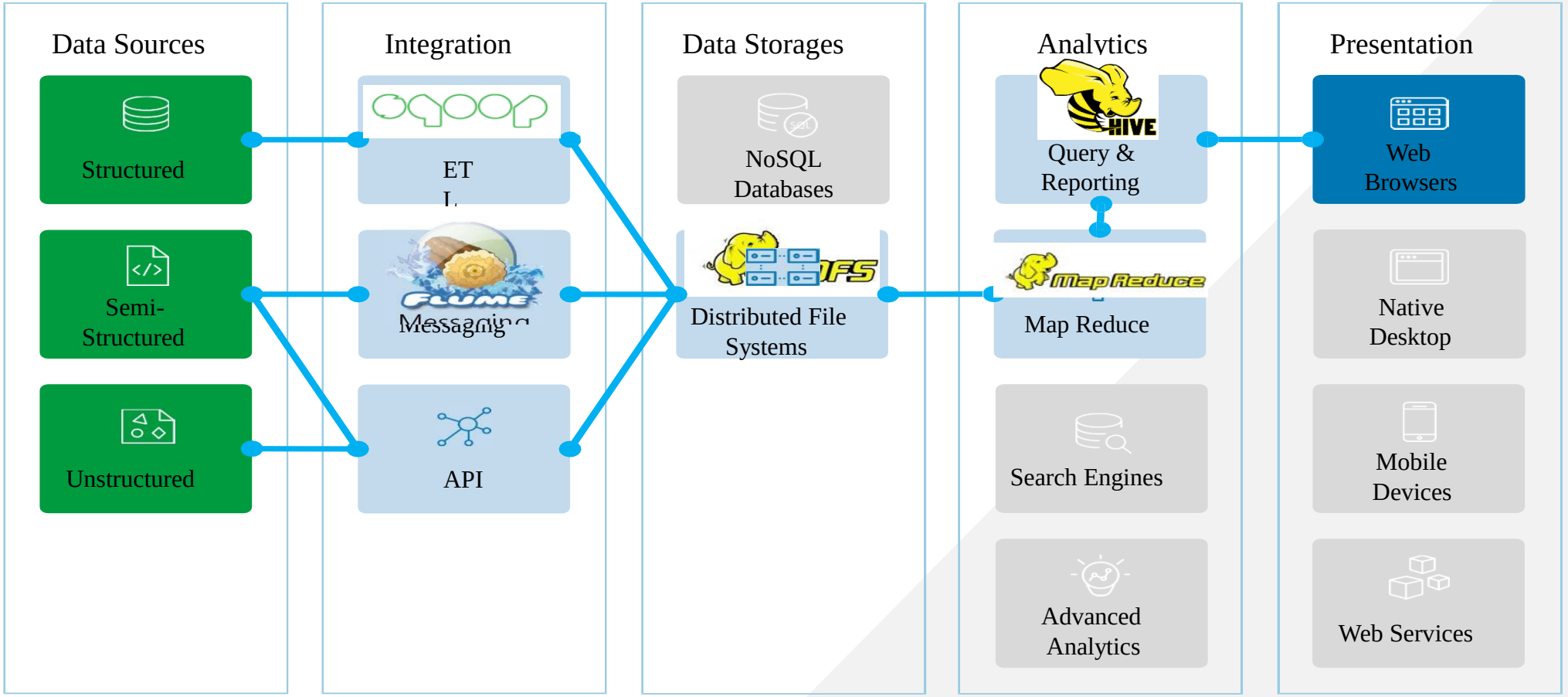


Key components introduced with non-relational movement

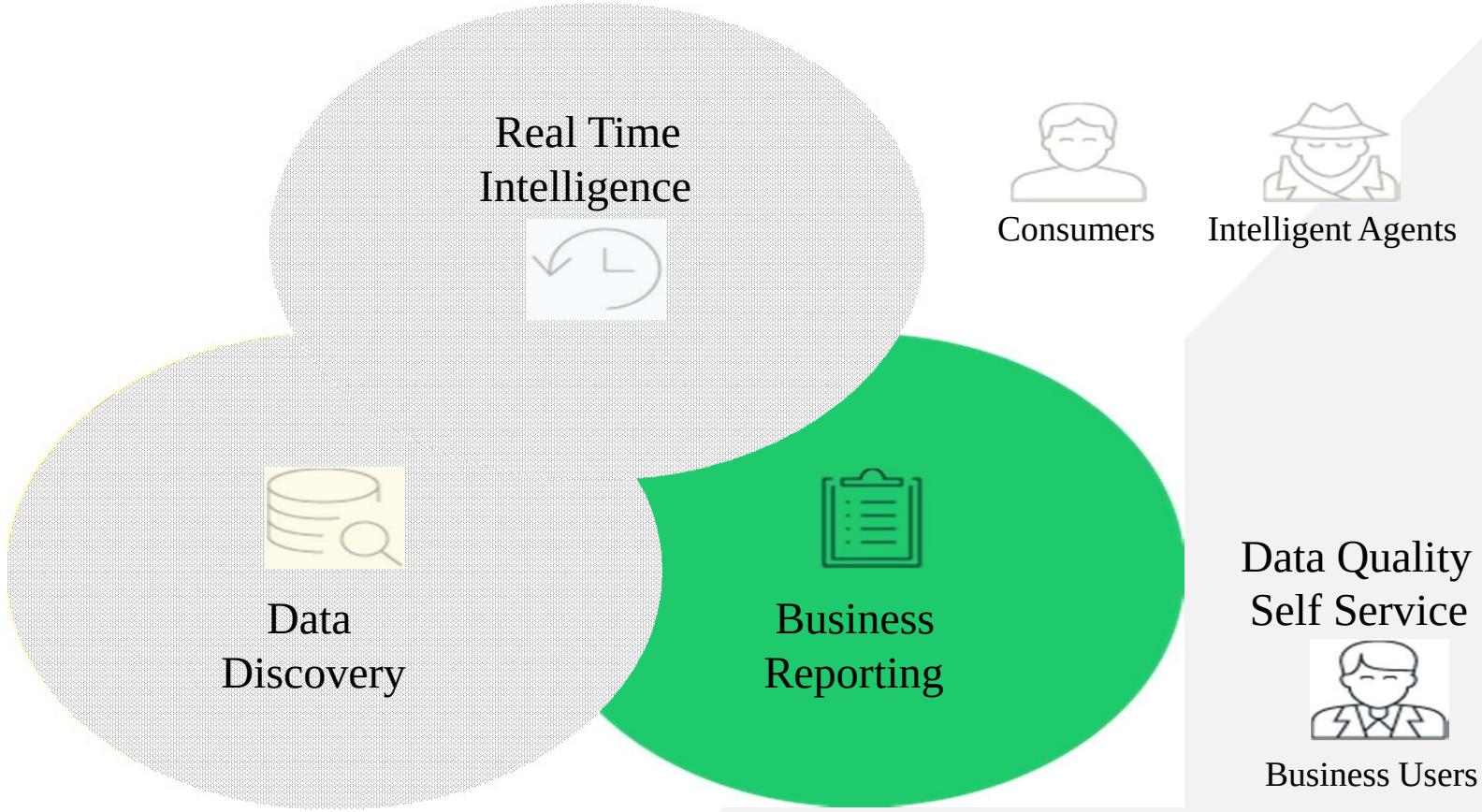
Big Data Analytics Use Cases



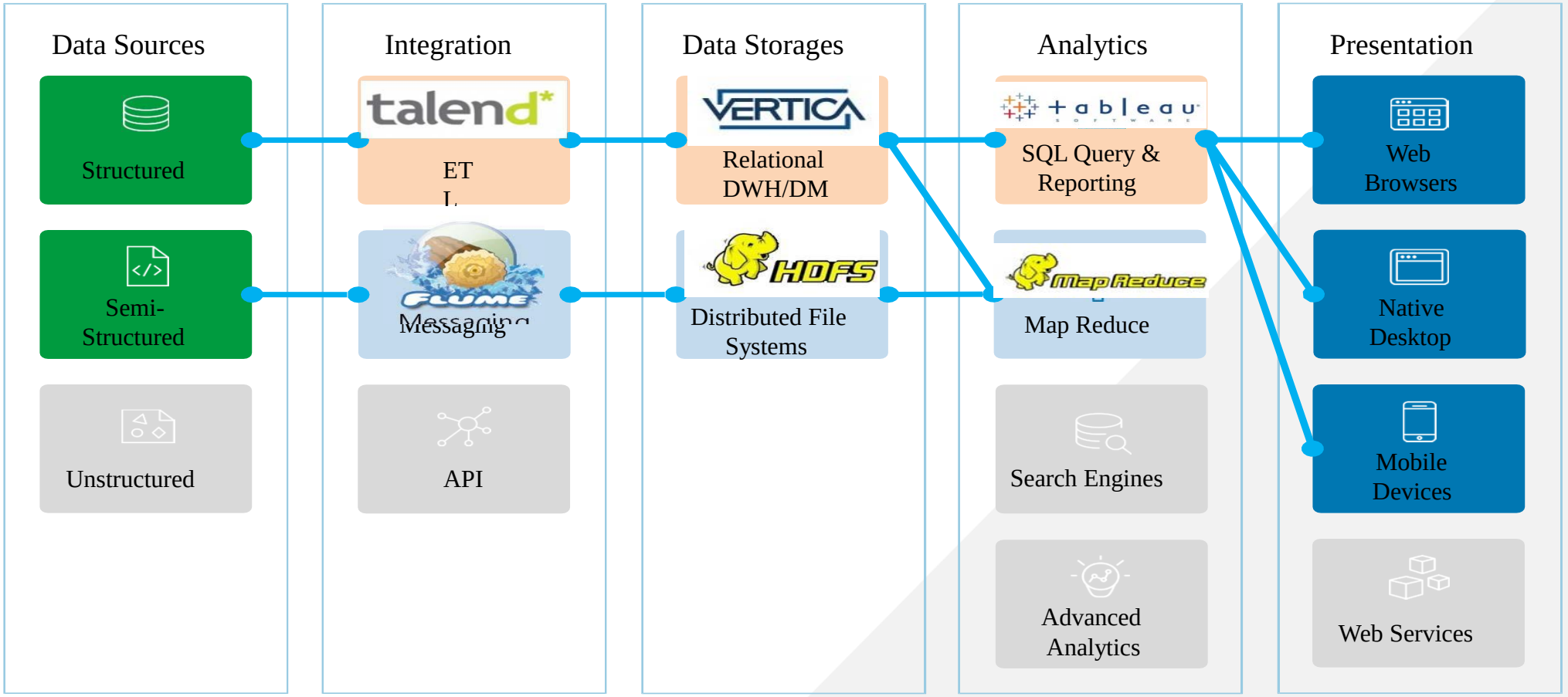
Data Discovery: Non-Relational Architecture



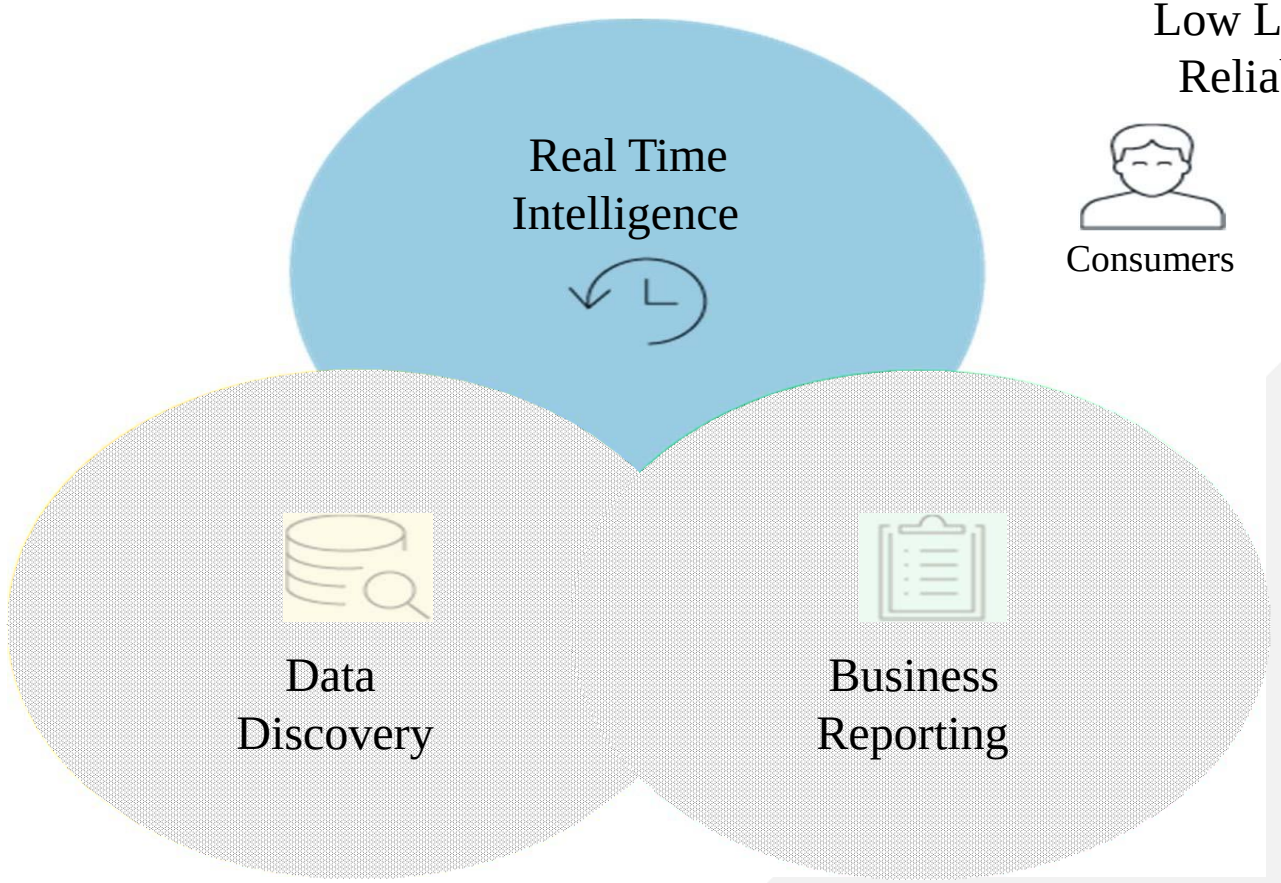
Big Data Analytics Use Cases



Business Reporting: Hybrid Architecture



Big Data Analytics Use Cases



Low Latency
Reliability

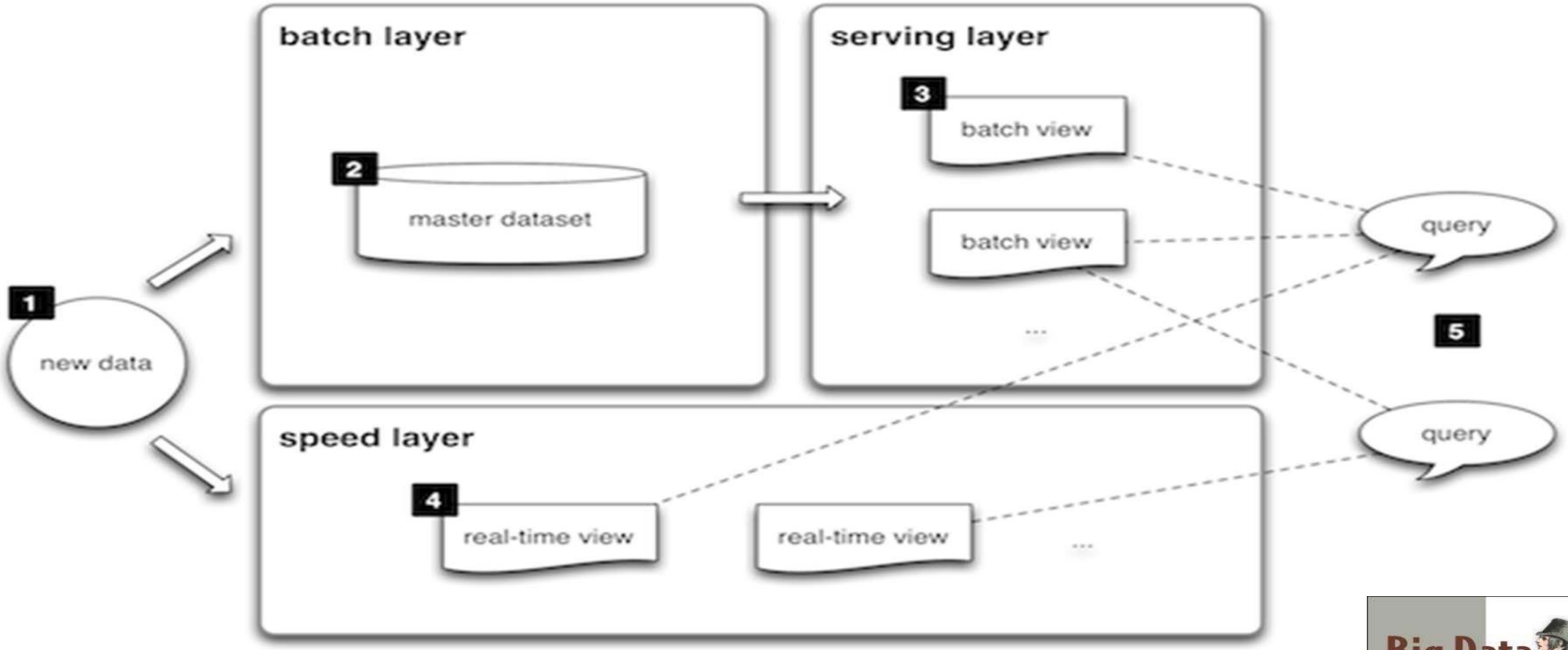
Consumers

Intelligent Agents

Data Scientists

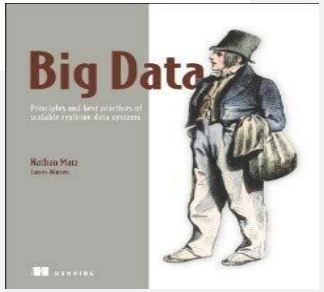
Business Users

Lambda Architecture



Lambda architecture is **a way of processing massive quantities of data (i.e. "Big Data") that provides access to batch-processing and stream-processing methods with a hybrid approach**.

Source:



Thank you

Please Send Your Queries on:

e-Mail: jitender.e14621@cumail.in