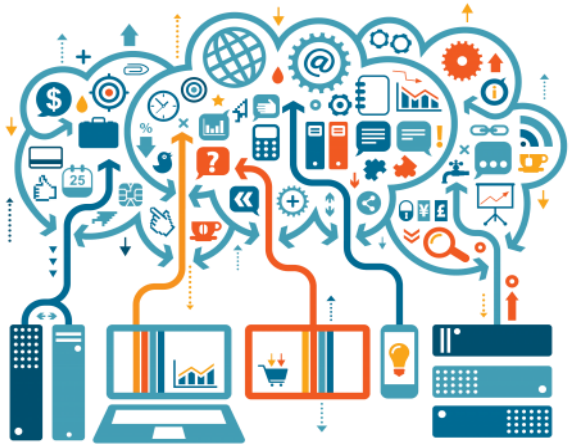
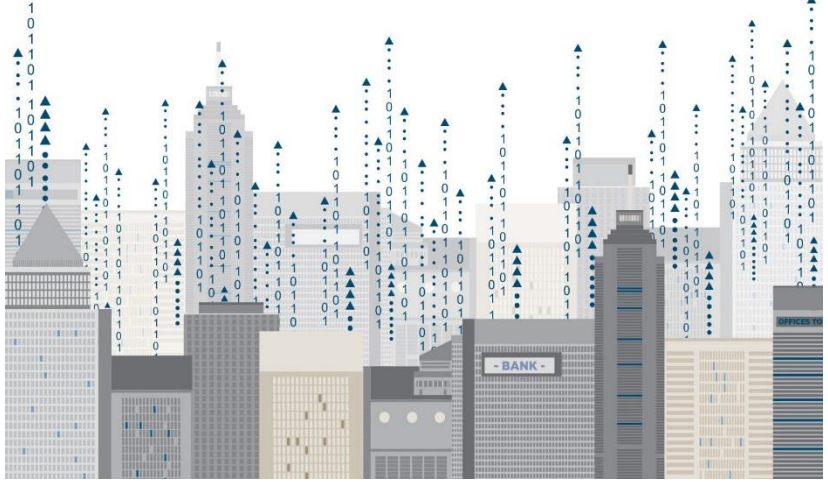




Big Data Analytics

Fall 2025

Lecture 1

Defining Big Data

- Extremely large data sets that have grown beyond the ability to manage and analyze with traditional data processing tools

- Conventional IT can no longer effectively handle either the size of the data set or the scale and growth of the data set

- Primary difficulties: acquisition, storage, searching, sharing, analytics and visualization

Gartner

- *John R. Mashey popularized the term*
- *Big data is high-volume, high-velocity and* high-variety *information assets that demand* cost-effective, innovative *forms of information processing for enhanced insight and decision making*. – Gartner

- https://cloud.google.com/learn/what-is-big-data
- https://en.wikipedia.org/wiki/Big_data

List of Global Big Data Businesses

- Energy
- Retail (Online/Brick and Mortar)
- Insurance
- Finance
- Actuarial
- Telco
- Healthcare
- Crime Prevention
- Governance
- Traffic
- Agricultural
- Media (Print/Electronic)
- Advertising
- Software houses
- Academic Institutions
- E-Commerce
- Consultancy Companies (Projects – Training – Workshops)
- Trading/Brokerage
- Oil and Gas
- Automotive

Big Data and Cloud Solution Companies

- Amazon AWS
- MS Azure
- Google
- HP
- Oracle
- Dell
- SAP
- SAS

- Apache (ASF)
- Cloudera
- Databricks
- Informatica
- Snowflake
- Confluent
- Redis Labs
- Matillion
- VMWare

- MongoDB
- REDIS
- Neo4J

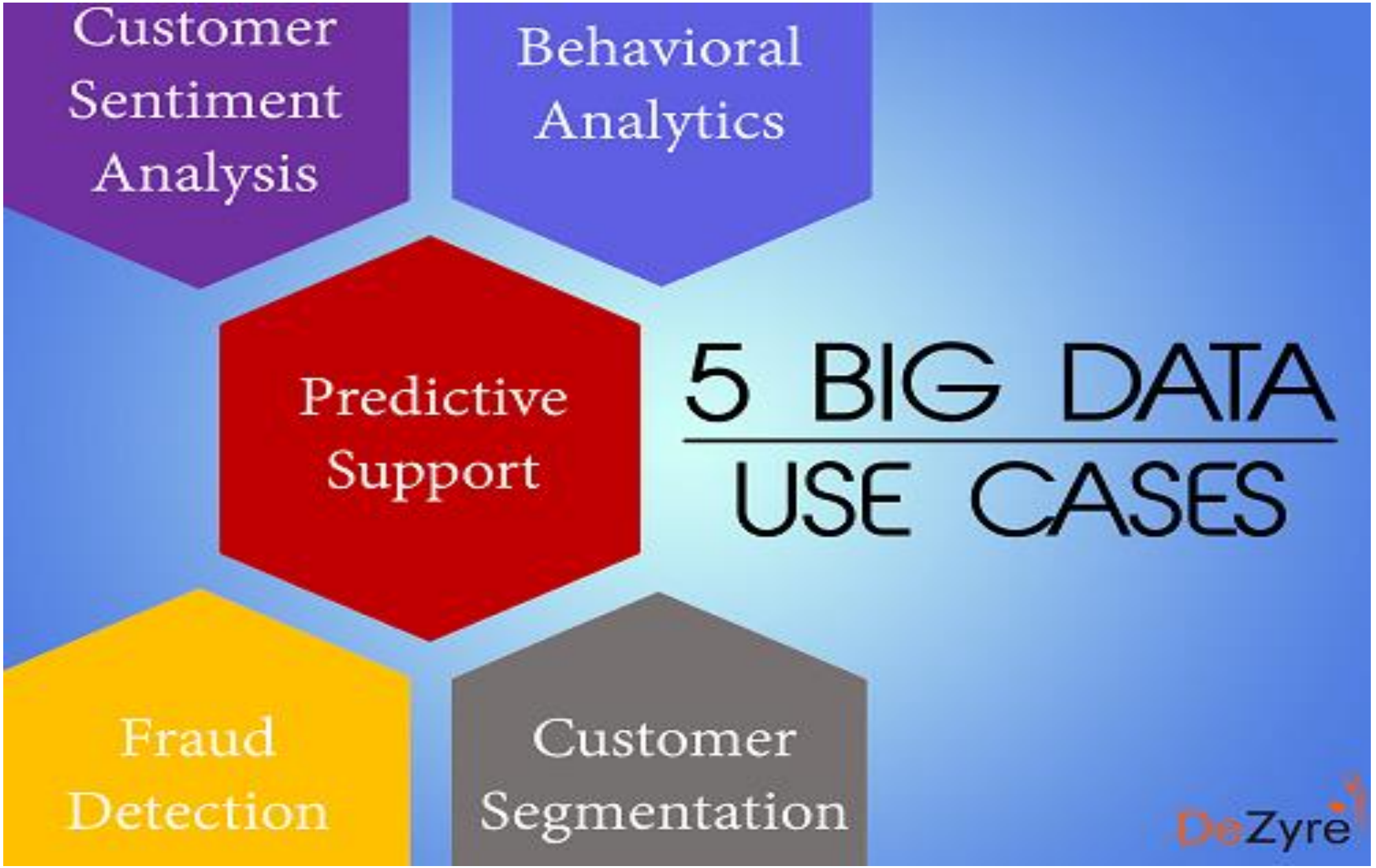
- Dell
- Intel
- Nvidia
- Cisco
- IBM

https://builtin.com/big-data/big-data-companies-roundup

https://themanifest.com/big-data/companies

https://db-engines.com/en/ranking


Volume
Size of Data
Validity
Data quality, Governace, Moster Data Management on Massive
Velocity
The Speed at which Data is Generated
Variability
Dynamic, Evolving Behavior in Data Source
BigData
Variety
Different type of Data
Venue
Distributed Heterogeneous Data from Multiple Platforms
Veracity
Data Accuracy
Vocabulary
Data Models, Semantics that describes data Structure
Value
Useful Data
Vagueness
Confusion over Meaning of BigData and Tools used


Customer Sentiment Analysis
Behavioral Analytics
Predictive Support
5 BIG DATA USE CASES
Fraud Detection
Customer Segmentation
DeZyre

BIG DATA USE CASES:

1. OPTIMIZE FUNNEL CONVERSION

2. BEHAVIORAL ANALYTICS

3. CUSTOMER SEGMENTATION

4. PREDICTIVE SUPPORT

5. MARKET BASKET ANALYSIS AND PRICING OPTIMIZATION

6. PREDICT SECURITY THREATS

7. FRAUD DETECTION

8. INDUSTRY SPECIFIC

Big Data opportunities across industries and use cases

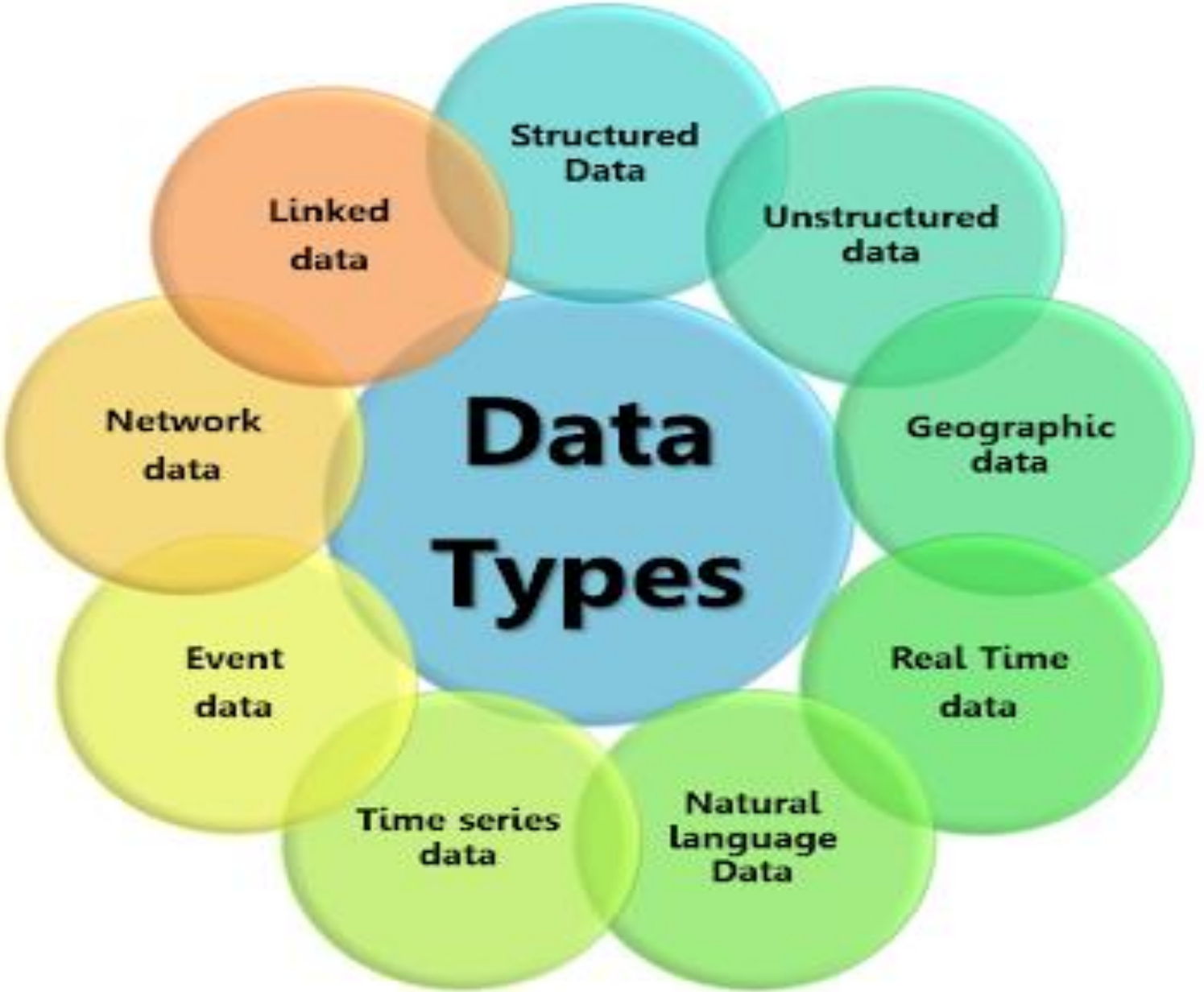
Innovative analytic use cases are cutting across structured, unstructured and semi structured data

Finance	Government	Telecom	Manufacturing	Energy	Healthcare
• Customer knowledge • Event marketing • Risk management	• Law enforcement • Video surveillance and security • Traffic flow optimization	• Analysis and customer retention • Analysis of network usage • Monitoring hearings and ad optimization	• Supply chain optimization • Defect tracking • RFID Correlation • Warranty management	• Weather forecasting • Natural resource exploration	• Drug development • Scientific research • Evidence based medicine • Healthcare outcomes

Horizontal use cases

- Sentiment analysis
- Social CRM / network analysis
- Churn mitigation
- Brand monitoring
- Cross and Up sell
- Loyalty & promotion analysis
- Web application optimization
- Marketing campaign optimization
- Brand management
- Social media analytics
- Pricing optimization
- Internal risk assessment
- Customer behavior analysis
- Revenue assurance
- Logistics optimization
- Clickstream analysis
- Influencer analysis
- IT infrastructure analysis
- Legal discovery
- Equipment monitoring
- Enterprise search

Sources: IDC: 2012 "Worldwide Big Data Technology and Services Forecast: 2011-2015, Gartner: 2012 "Big Data Drives Rapid Changes in Infrastructure and $232 Billion in IT Spending Through 2016


Data Types
Structured Data
Unstructured data
Geographic data
Real Time data
Natural language Data
Time series data
Event data
Network data
Linked data

Unstructured Data Types for Big Data Analysis

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

Why BDA can fail?

- Asking the wrong questions and solving the wrong problem
- Lack of right data
- Lack of right talent
- Not deploying products – we think deployment is the last step (DevOps)
- Not involving the business
- Selection of the wrong technology stack – lack of resource optimization
- Indifference to technological requirements

https://www.kdnuggets.com/2021/09/sparkbeyond-avoid-data-science-projects-fail.html

https://medium.com/@daniel_3607/addressing-the-85-data-project-failure-rate-is-your-companys-greatest-chance-to-succeed-da37967372d6

https://www.globalbigdataconference.com/news/141927/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed.html

https://hbr.org/2020/02/use-this-framework-to-predict-the-success-of-your-big-data-project

https://www.datascience-pm.com/project-failures/

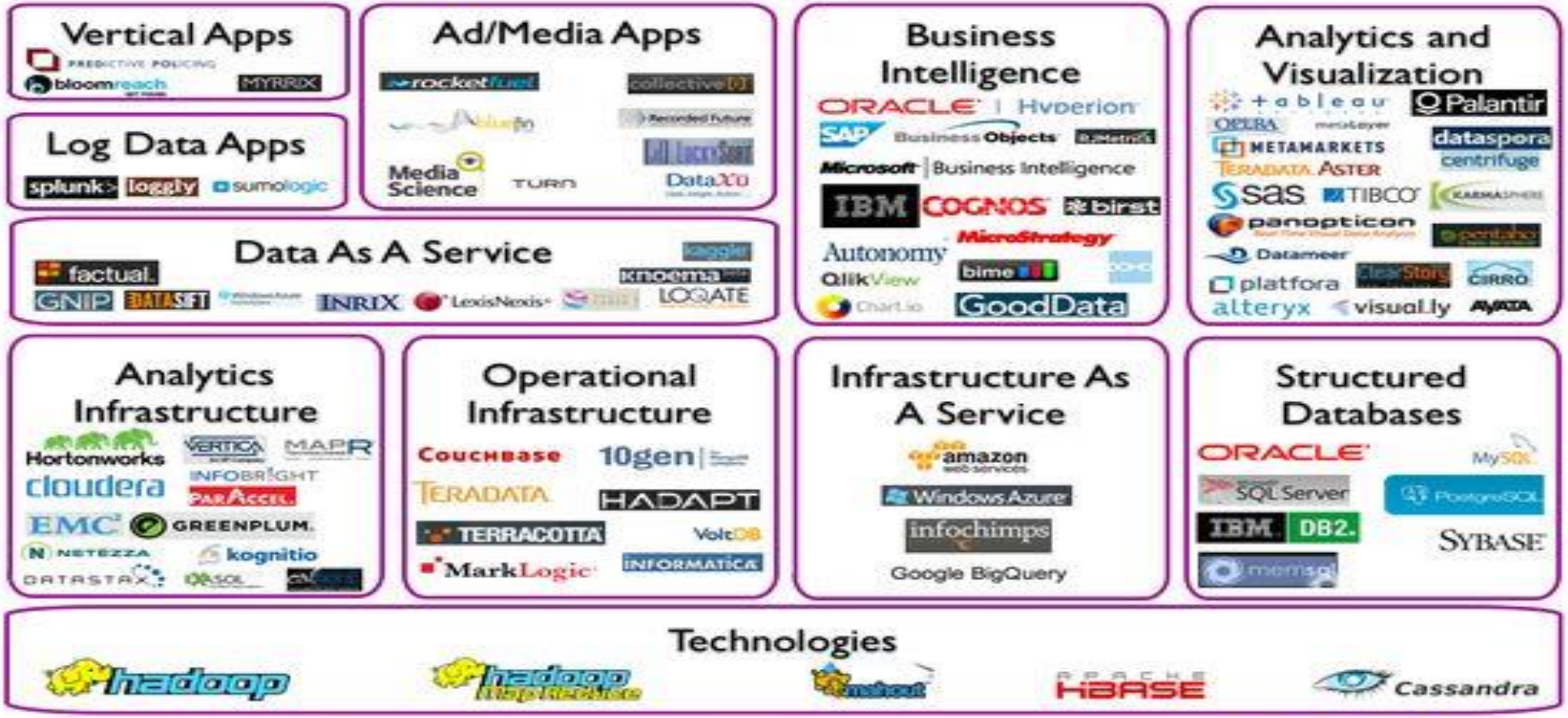
Disruption

- Not Disruptive Now – Not Scary Anymore
  - Due to cloud solutions
  - If you fail, it is not anymore due to hardware or software limitations
- BDA Project Generic Process:
  - Acquire understanding of business operations
  - Understand the big data business problem
  - Do a benefit analysis – Will BDA be aligned with goals? Will there be actual insights? Will they be actionable? What are short and long term benefits?
  - Measure risk: What's the loss if you didn't do it? - Resolution of Pain Points
  - Determine previous successful use cases
  - Review BDA Implementation Options – Technology stack is cloud now
  - Implement

Big Data in a state of Flux: Evolution of Tools, Technologies and Procedures

However, Orgs Cannot Wait-and-See!

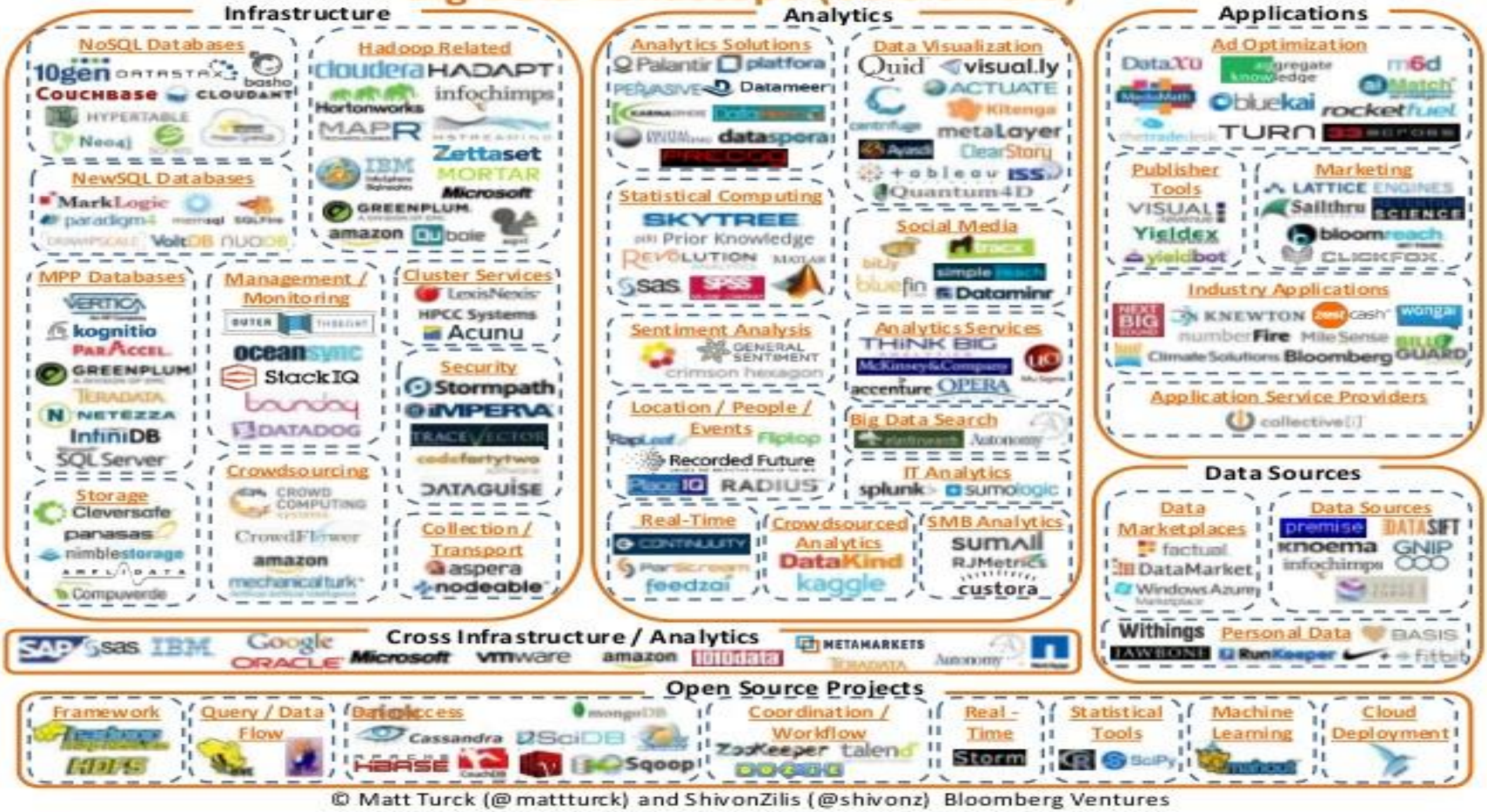
Big Data Landscape



dave@vcdave.com

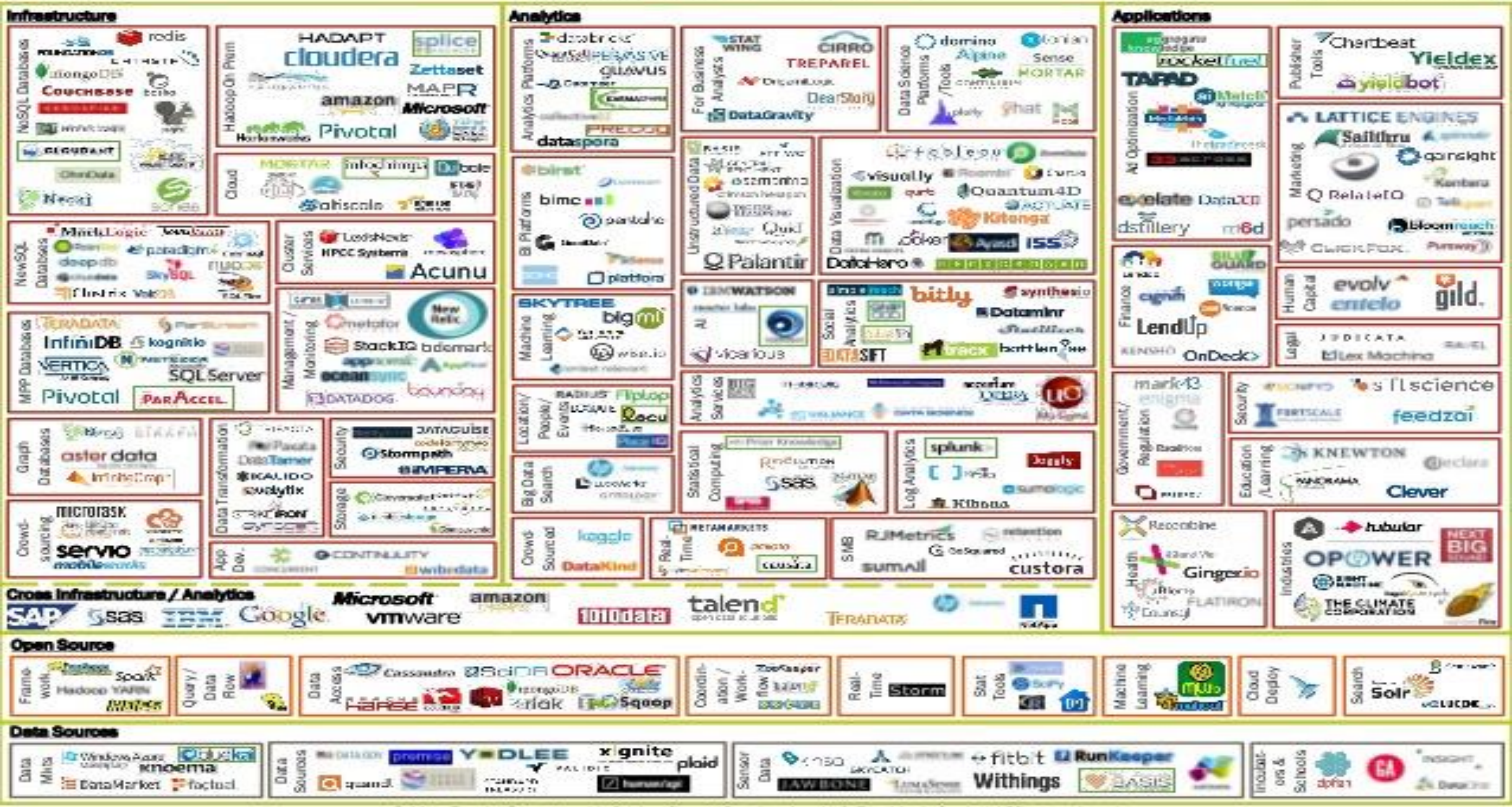
blogs.forbes.com/davefeinleib

Big Data Landscape (Version 2.0)

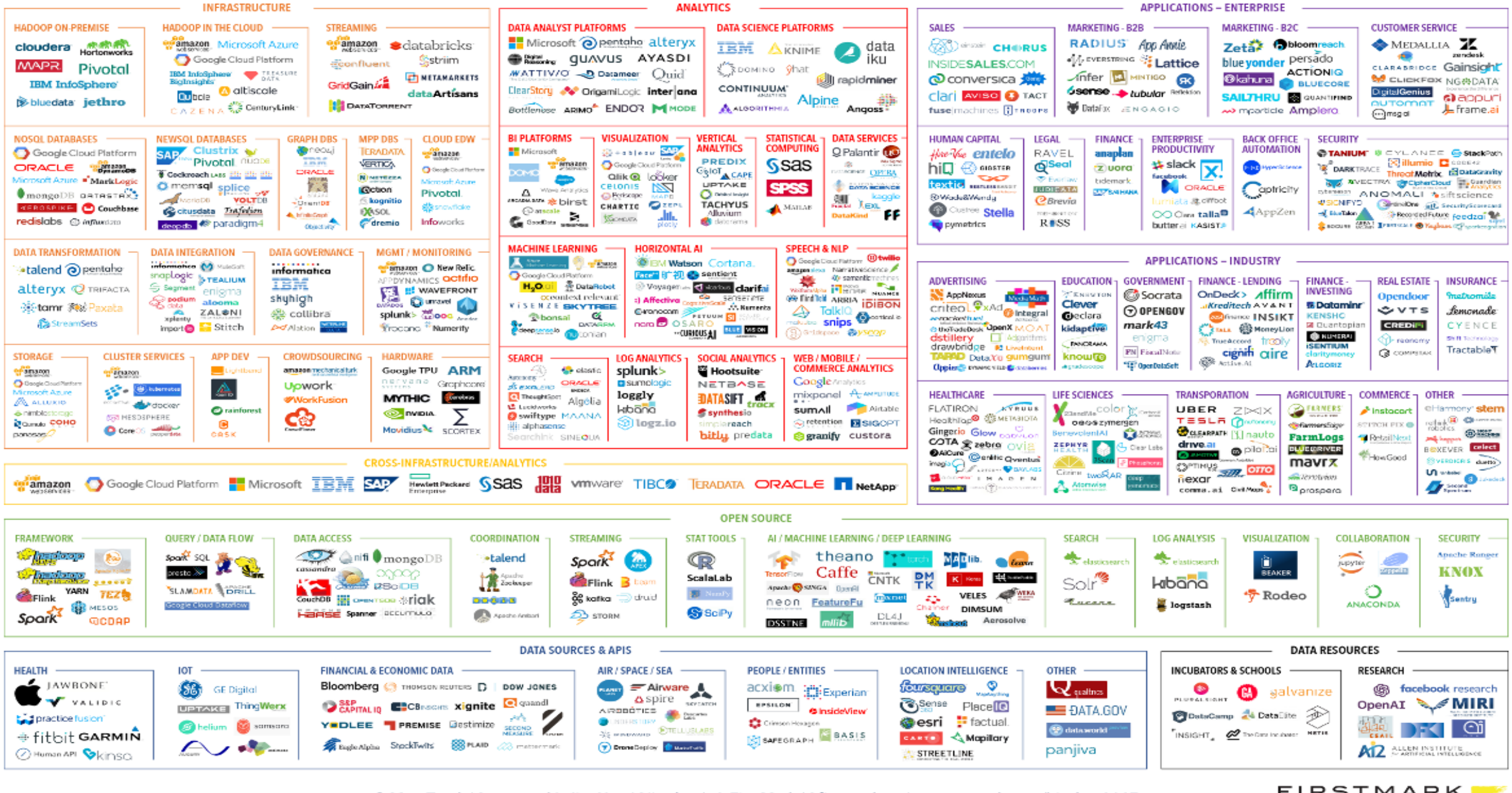


BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO



BIG DATA LANDSCAPE 2017

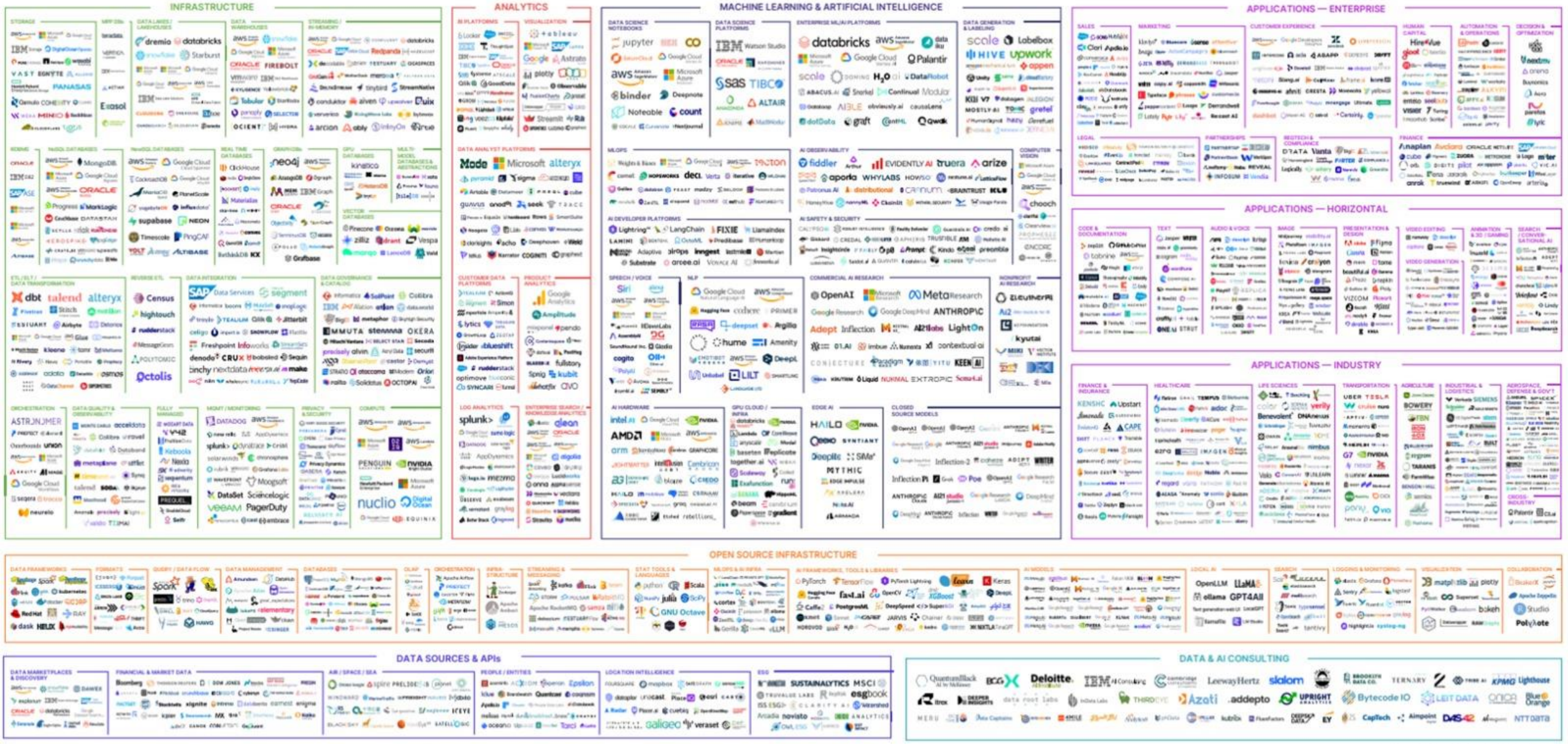


Last updated 4/5/2017

mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

https://mad.firstmark.com/

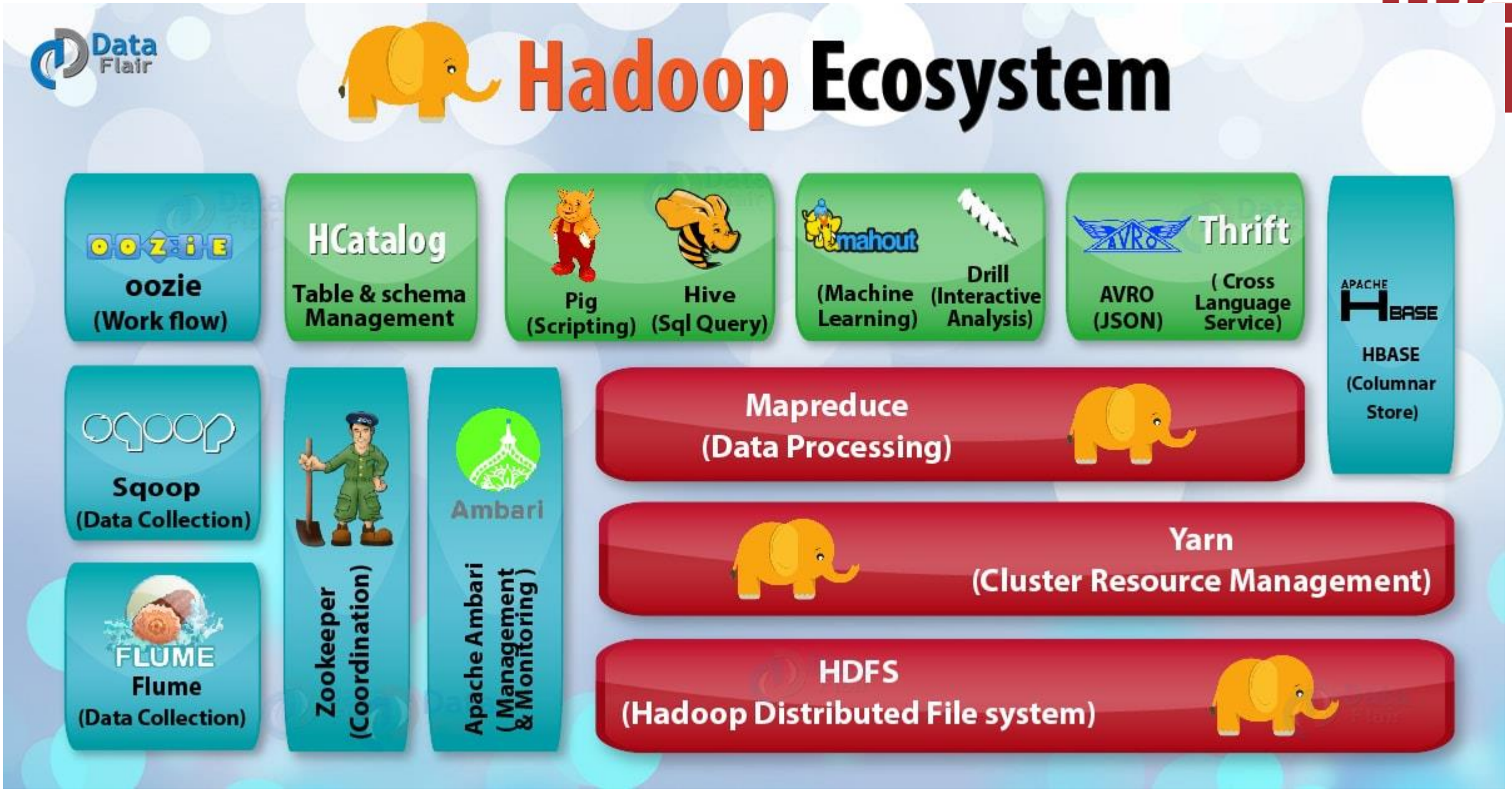


Version 1.0 - March 2024 © Matt Turck (@mattturck) , Aman Kabeer (@AmanKabeer11) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2024 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2024@firstmarkcap.com

FIRSTMARK EARLY STAGE VENTURE CAPITAL

Hadoop

- Hadoop has been around some time (Google MapReduce and Big Table, Doug Cutting, Hadoop, Yahoo): – works well for tasks which are deep and computationally extensive and no time restriction (clustering)
- Removes management overhead
- Fault Tolerant
- Commodity Hardware
- Different Configs for Different Hardware – a blend!
- Network is a bottleneck in BDA so Hadoop works where the data resides, hence minimizing communication


Data Flair
Hadoop Ecosystem
oozie
(Work flow)
HCatalog
Table & schema Management
Pig
(Scripting)
Hive
(Sql Query)
mahout
(Machine Learning)
Drill
(Interactive Analysis)
AVRO
(JSON)
Thrift
(Cross Language Service)
APACHE HBASE
HBASE
(Columnar Store)
Sqoop
(Data Collection)
Zookeeper
(Coordination)
Ambari
Apache Ambari
(Management & Monitoring)
Mapreduce
(Data Processing)
FLUME
Flume
(Data Collection)
Yarn
(Cluster Resource Management)
HDFS
(Hadoop Distributed File system)

Hadoop Success Stories

- https://www.forbes.com/sites/bernardmarr/2015/11/16/big-data-success-stories-beyond-hadoop/
- https://www.singlestore.com/blog/stampeding-away-from-hadoop-three-customer-success-stories/
- https://ellicium.com/success-stories/managed-services-for-hadoop/
- https://www.bmc.com/blogs/hadoop-examples/
- https://community.microstrategy.com/s/article/Customer-Success-Stories-with-the-Hadoop-Gateway?language=en_US

Note!

- Not a plug and play experience
- Skill set, experience, technical requirements, business decisions
- HDFS and HBase for Data management
- MapReduce / Spark / Oozie as a Processing framework
- Hive as Development framework
- Tableau / PowerBI / SiSense / Microstrategy for BI

Skillset Needed

- Data Science: Ability to translate the significance of data in a way that can be easily understood by others
- Machine Learning/Statistics/Linear Algebra: Important to understand the basic mathematics behind each algorithm
- Data Cleaning: Wrangling, Preprocessing, Dredging
- Data Analysis experience: SQL-based, DWH-based, BI-based
- Data Visualization (BI tools and Python-based dashboards)
- Data Discovery: Experience in Extraction of Insights
- Skill in Linux OS
- Skill in Python development using open source technologies

Big Data Techniques

- Parallel Processing, Distributed Computing and High-Performance Computing
- Motherboard, DRAM, Cache, SSD/HDD, GPU)
- In-memory processing, columnar SQL access
- NoSQL Databases
- Containerization (Docker)
- Infrastructures and architectures (lambda, kappa, htap)
- Cloud Computing
- High-Speed Networks
- AI and Data Science

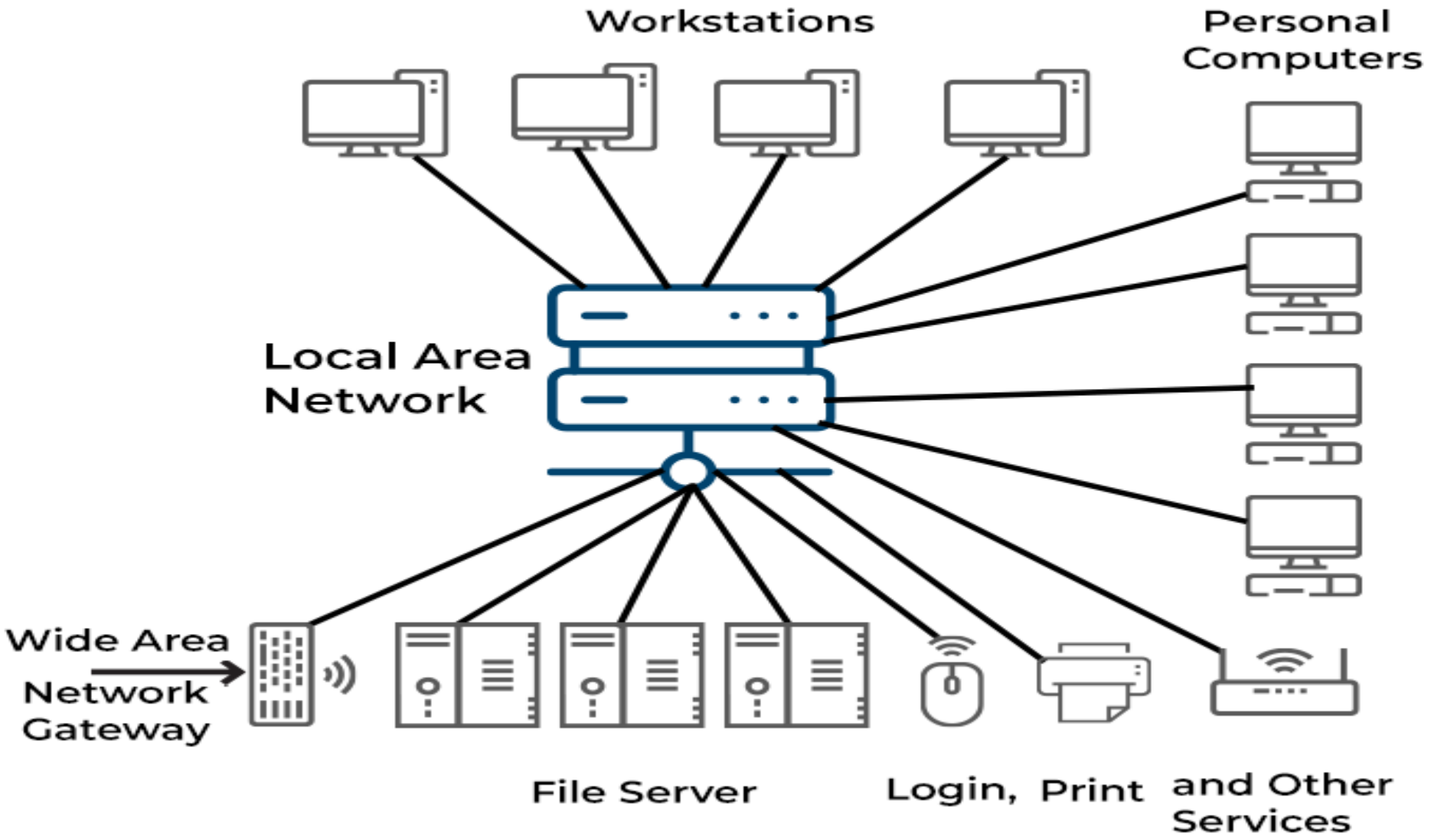
:)

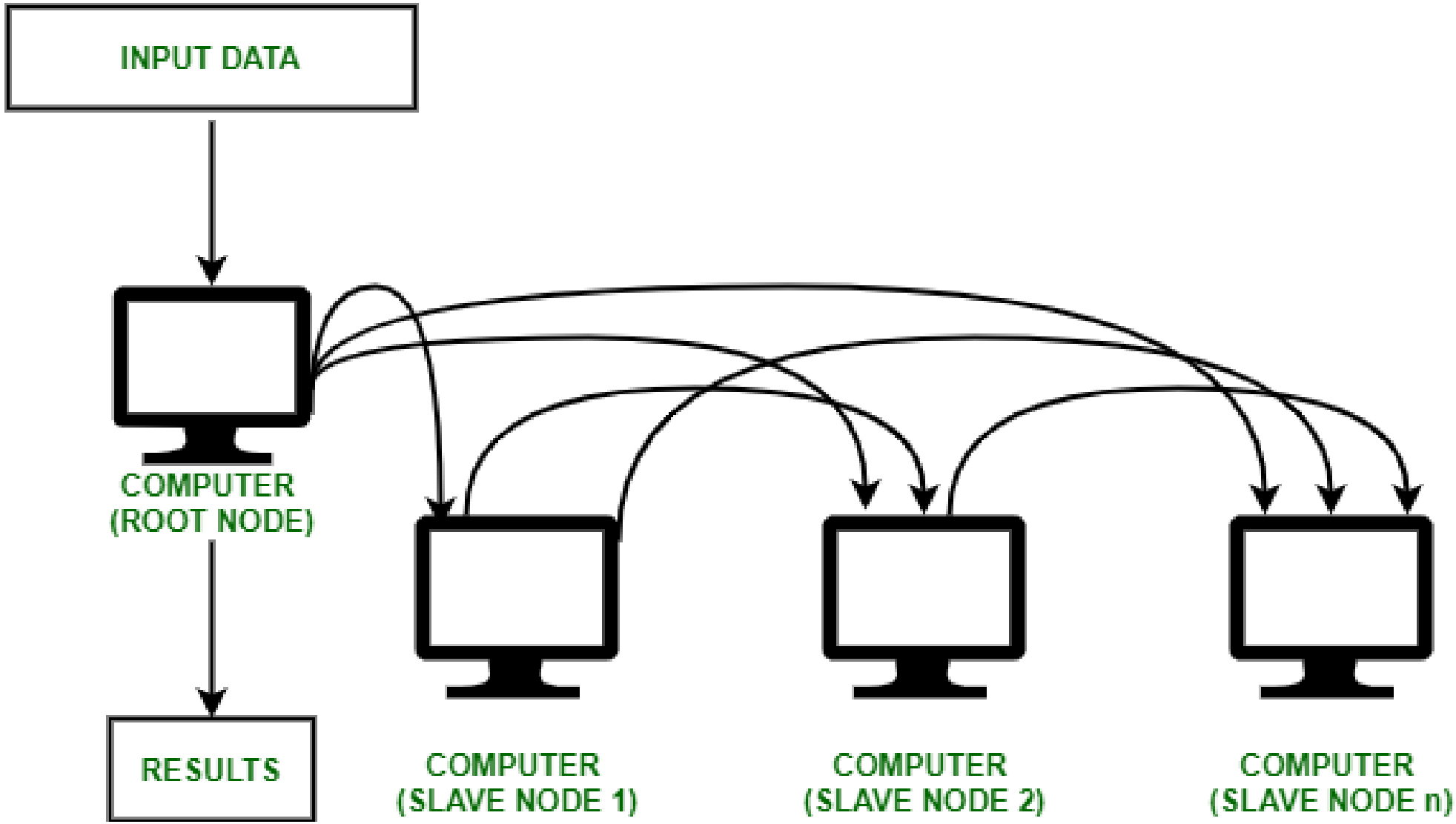
Yes, parallel processing is used in normal Windows operations on a Core i7 processor. Modern operating systems like Windows are designed to take advantage of multi-core processors, such as the Intel Core i7, to improve performance and responsiveness. Here's how parallel processing is utilized:

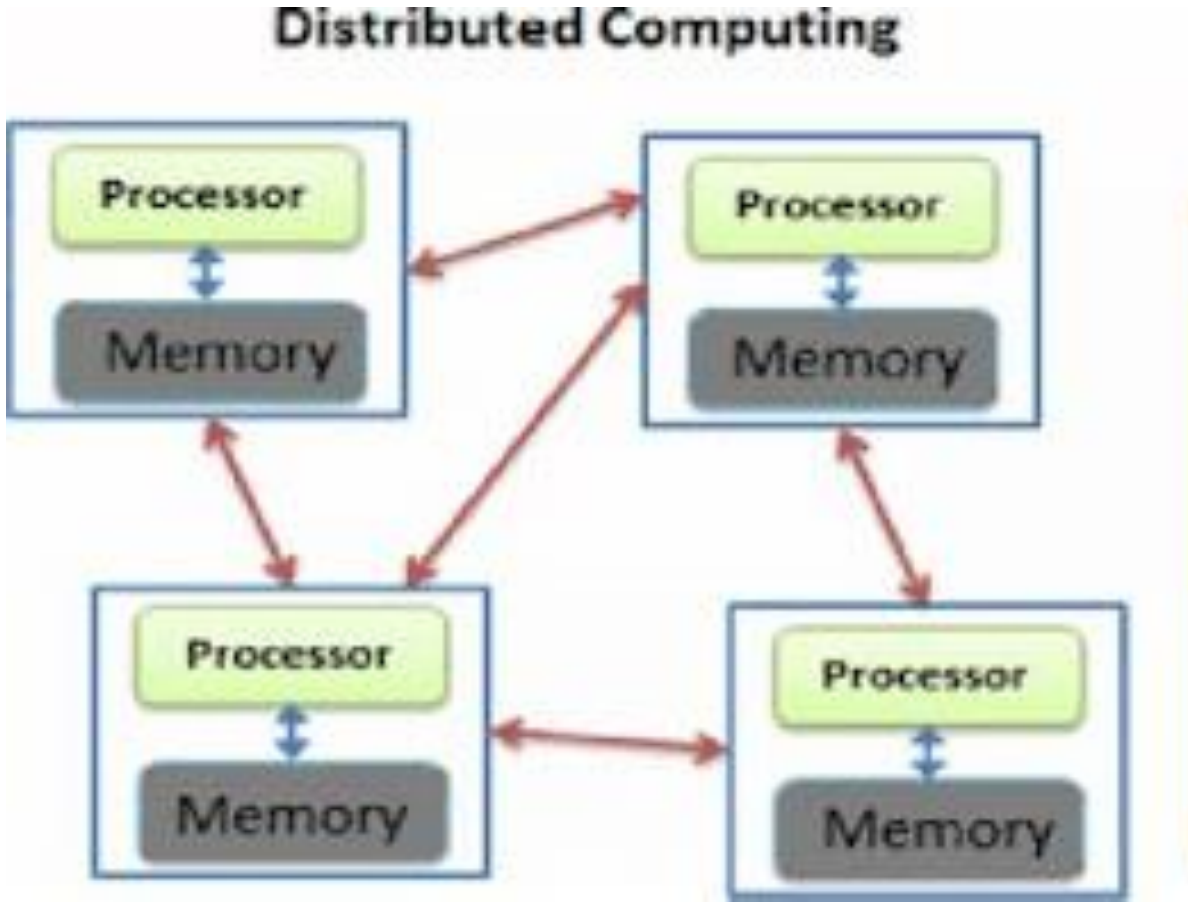
1. **Multithreading**: Many applications, including the Windows operating system itself, are designed to run multiple threads simultaneously. Each thread can be processed by a different core, allowing tasks to be completed more quickly.
2. **Task Scheduling**: The Windows operating system includes a task scheduler that efficiently distributes processes across the available cores. This helps in balancing the load and ensuring that no single core is overwhelmed, leading to smoother performance.
3. **Background Processes**: Windows often runs several background processes concurrently, such as system maintenance tasks, updates, and security scans. Parallel processing allows these tasks to run without significantly impacting the performance of the main tasks or applications you are using.
4. **Application Performance**: Many modern applications, especially those that are resource-intensive like video editing software, games, and scientific computing applications, are optimized for multi-core processors. They can split their workload across multiple cores to run more efficiently.

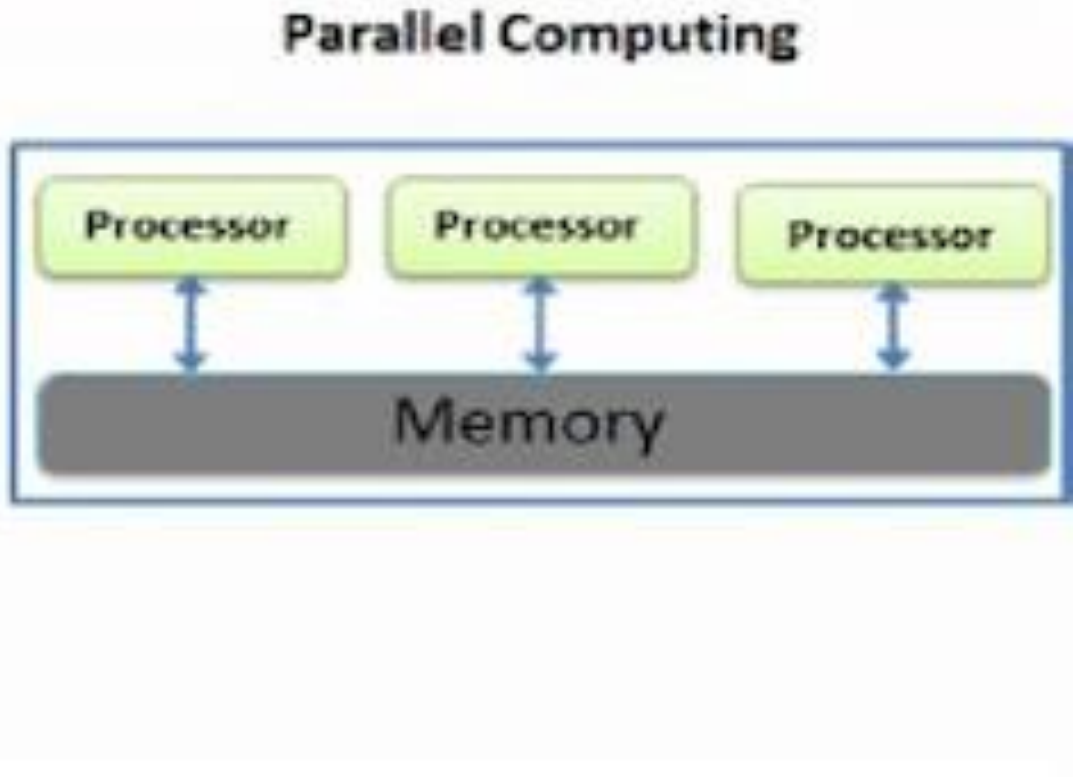
TOOLBOX™

A DISTRIBUTED SYSTEM




INPUT DATA
COMPUTER
(ROOT NODE)
RESULTS
COMPUTER
(SLAVE NODE 1)
COMPUTER
(SLAVE NODE 2)
COMPUTER
(SLAVE NODE n)


Distributed Computing
Processor
Memory
Processor
Memory
Processor
Memory
Processor
Memory


Parallel Computing
Processor
Processor
Processor
Memory

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:

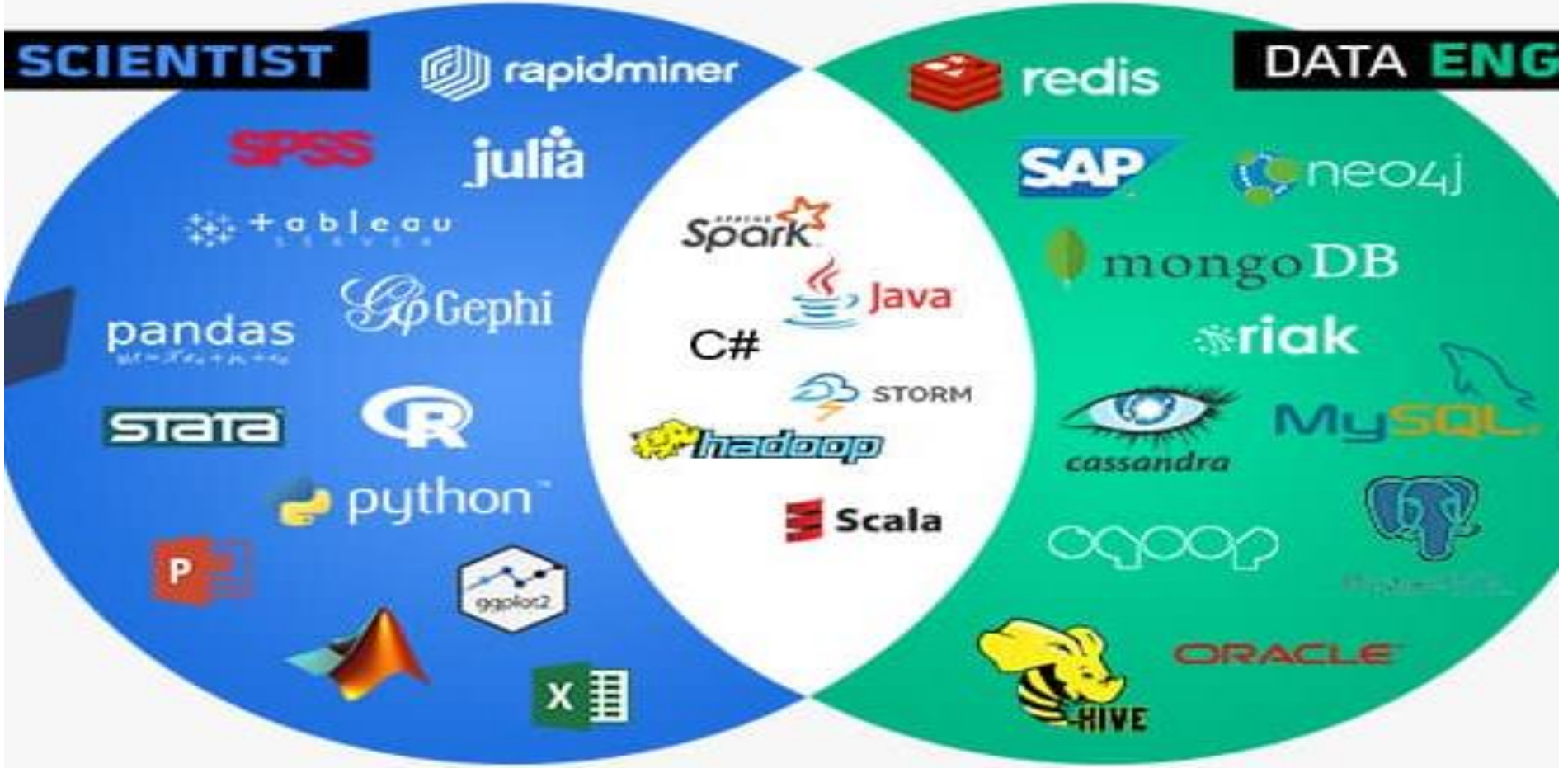
Excel, Tableau, SQL

Data Engineer vs Data Scientist

Data Engineer	Data Scientist
Programming Languages like R, Python	Programming Languages like R, Python
Database management skills including SQL	Math, Statistics
Knowledge of big data architecture such as Hadoop, Kafka, Hive, Impala	Knowledge of database queries to access data
Knowledge of security protocols to ensure data is secure.	Machine learning skills.
Cloud computing skills, e.g., Amazon web services.	Communication skills
Data Visualization skills.	Data Visualization skills.
Basics of machine learning added benefit.	NLP, AI added benefit.

ProjectPro

LANGUAGES, TOOLS AND SOFTWARE

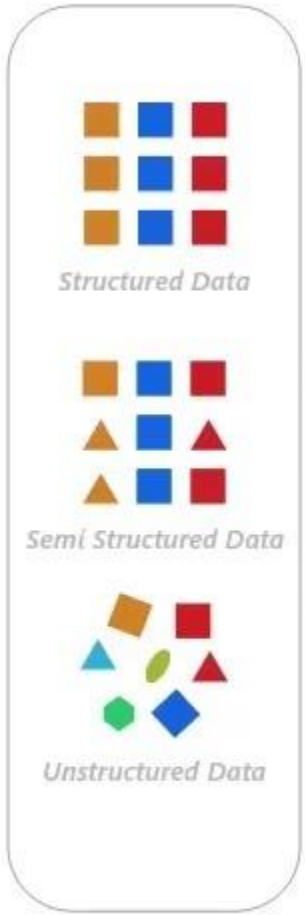


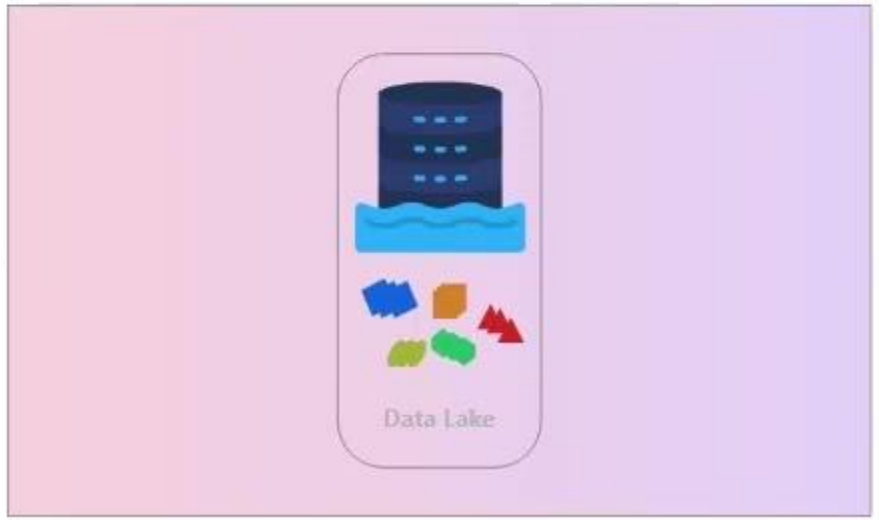
Big Data Analytics (BDA)

- Big Data Analytics ≡ Derive Value ≡ Strike Gold!
- Learns models which give you information that is hidden, non-trivial, critical, previously unknown and potentially useful (valuable)
- Benefits: Drive business, formulate business strategies, and maintain competitive edge.

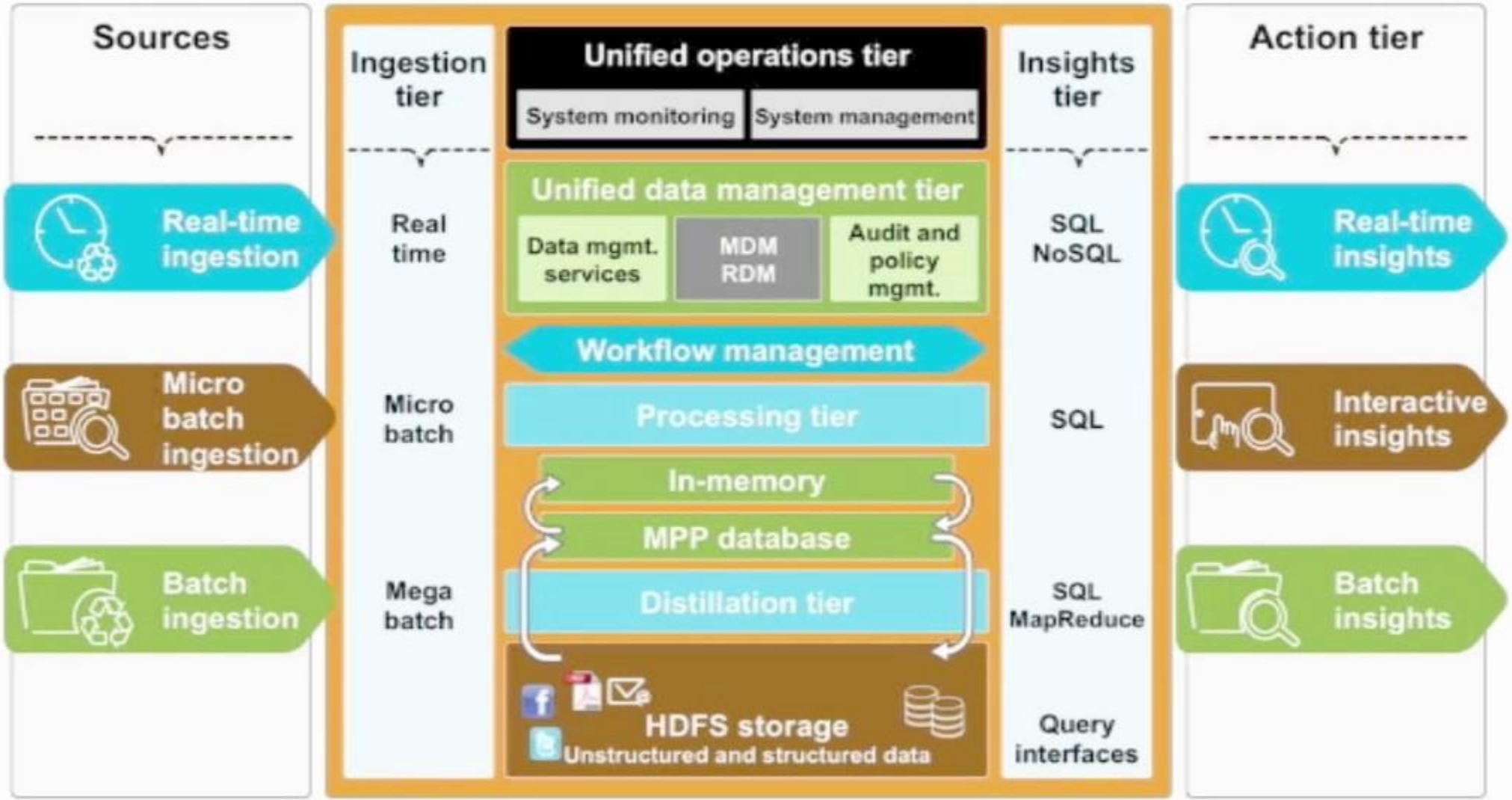
Big Data Workflow

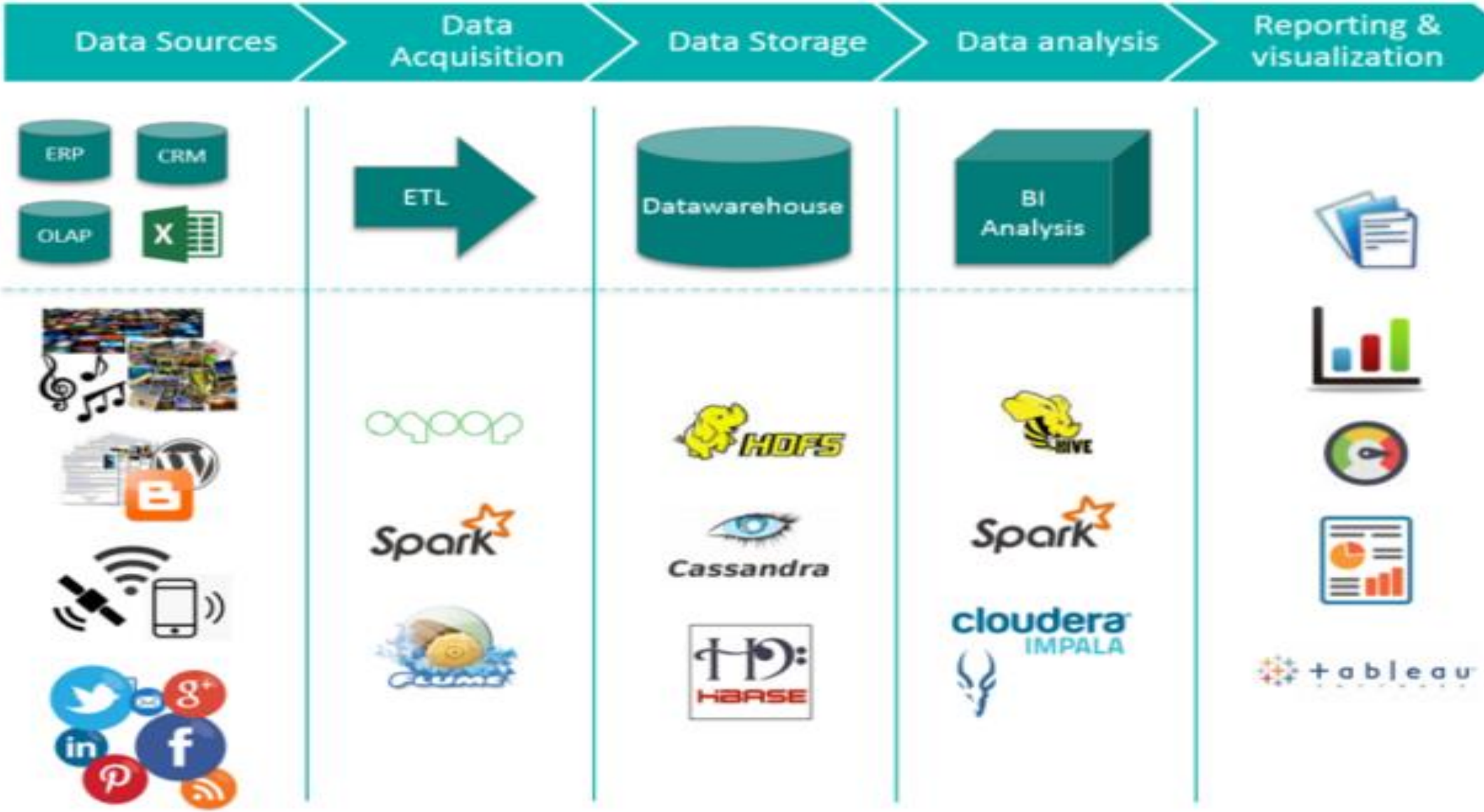
RadioStudio





Data Lake


Sources
Real-time ingestion
Micro batch ingestion
Batch ingestion
Ingestion tier
Real time
Micro batch
Mega batch
Unified operations tier
System monitoring
System management
Unified data management tier
Data mgmt. services
MDM RDM
Audit and policy mgmt.
Workflow management
Processing tier
In-memory
MPP database
Distillation tier
HDFS storage
Unstructured and structured data
Insights tier
SQL NoSQL
SQL
SQL MapReduce
Query interfaces
Action tier
Real-time insights
Interactive insights
Batch insights


Data Sources
Data Acquisition
Data Storage
Data analysis
Reporting & visualization
ERP
CRM
OLAP
X
ETL
Datawarehouse
BI Analysis
oozie
Spark
Flume
HDFS
Cassandra
HBASE
HIVE
Spark
cloudera
IMPALA
tableau

BDA Situation

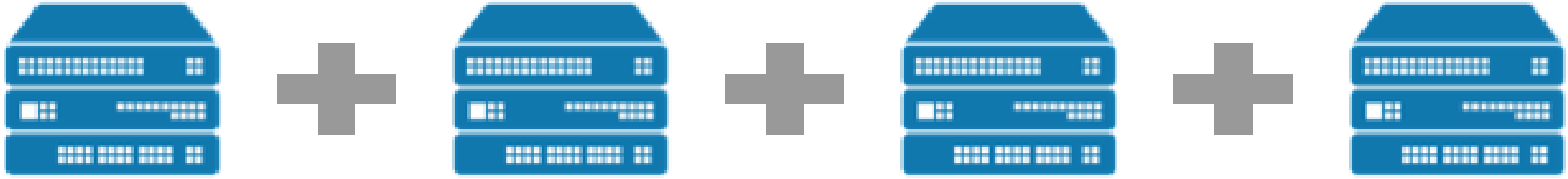
Era	Tools	Data Focus	Output	Role of AI/LLMs
Classic BDA (2010s)	Hadoop, Hive, Spark	Structured/semi-structured	Batch reports, dashboards	Minimal
Modern BDA (2020s)	Cloud DW (Snowflake, BigQuery), Spark Streaming	Structured + some unstructured	Real-time dashboards, ML predictions	ML models integrated
BDA + LLMs (2023–2025)	LLMs, Vector DBs, RAG, AI copilots	Structured + **unstructured (text, image, audio)**	Conversational insights, predictive + prescriptive analytics	LLMs generate queries, automate wrangling, summarize, recommend

BDA Situation

Era	Architecture	Tools & Tech	Data Focus	Output	Role of AI/LLMs
Classic BDA (2010s)	**Data Lakes (HDFS-based)** → centralized storage of raw data	Hadoop, Hive, Spark, Pig	Mostly **structured/semi-structured** (logs, tables, JSON)	Batch analytics, ETL to warehouse, static dashboards	Minimal → manual SQL/HQL
Modern BDA (2020s)	**Cloud Data Lakes + Early Data Fabric** (integration across sources)	Spark Streaming, Snowflake, BigQuery, Lakehouse (Databricks Delta Lake)	Structured + **some unstructured** (images, IoT streams)	Real-time dashboards, ML predictions	ML integrated into pipelines, some automation
BDA + LLMs (2023–2025)	**Intelligent Data Fabric over Data Lakes** (AI-driven metadata mgmt, knowledge graphs, vector DBs)	LLMs, RAG, Vector DBs (Pinecone, FAISS, Milvus), Databricks MosaicML, Snowflake Cortex	**All data types**: structured + **unstructured (text, docs, video, audio, sensor data)**	Conversational insights, predictive + prescriptive analytics, autonomous actions	LLMs power natural language queries, automate wrangling, augment decision-making, bridge data silos

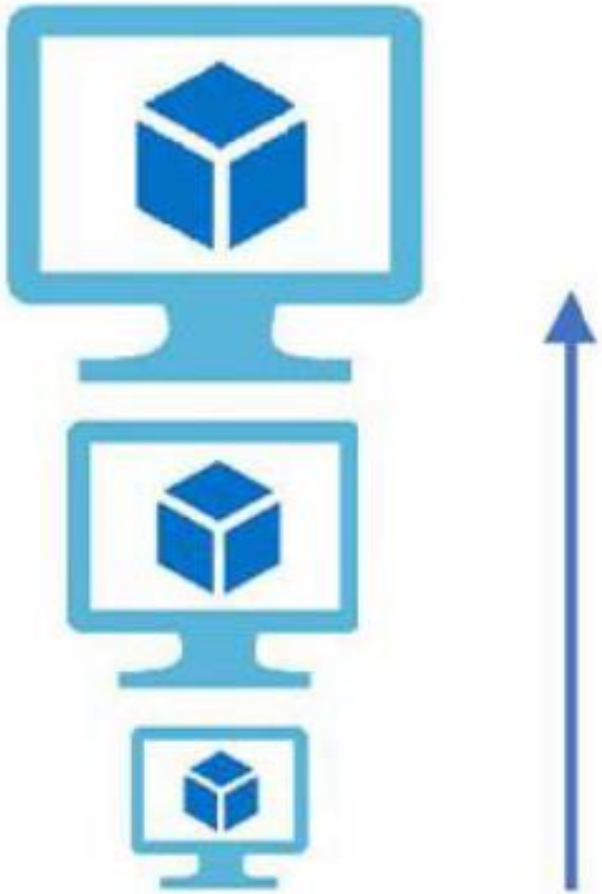
Scale Up

Scale Out



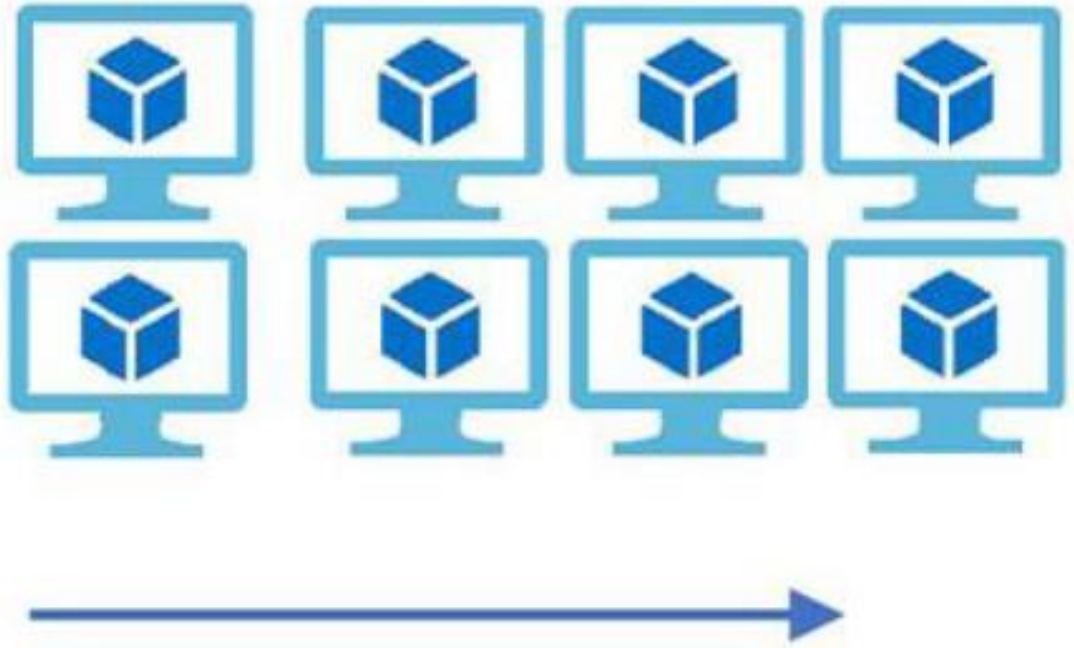
Vertical Scaling

(Increase size of instance (RAM , CPU etc.))

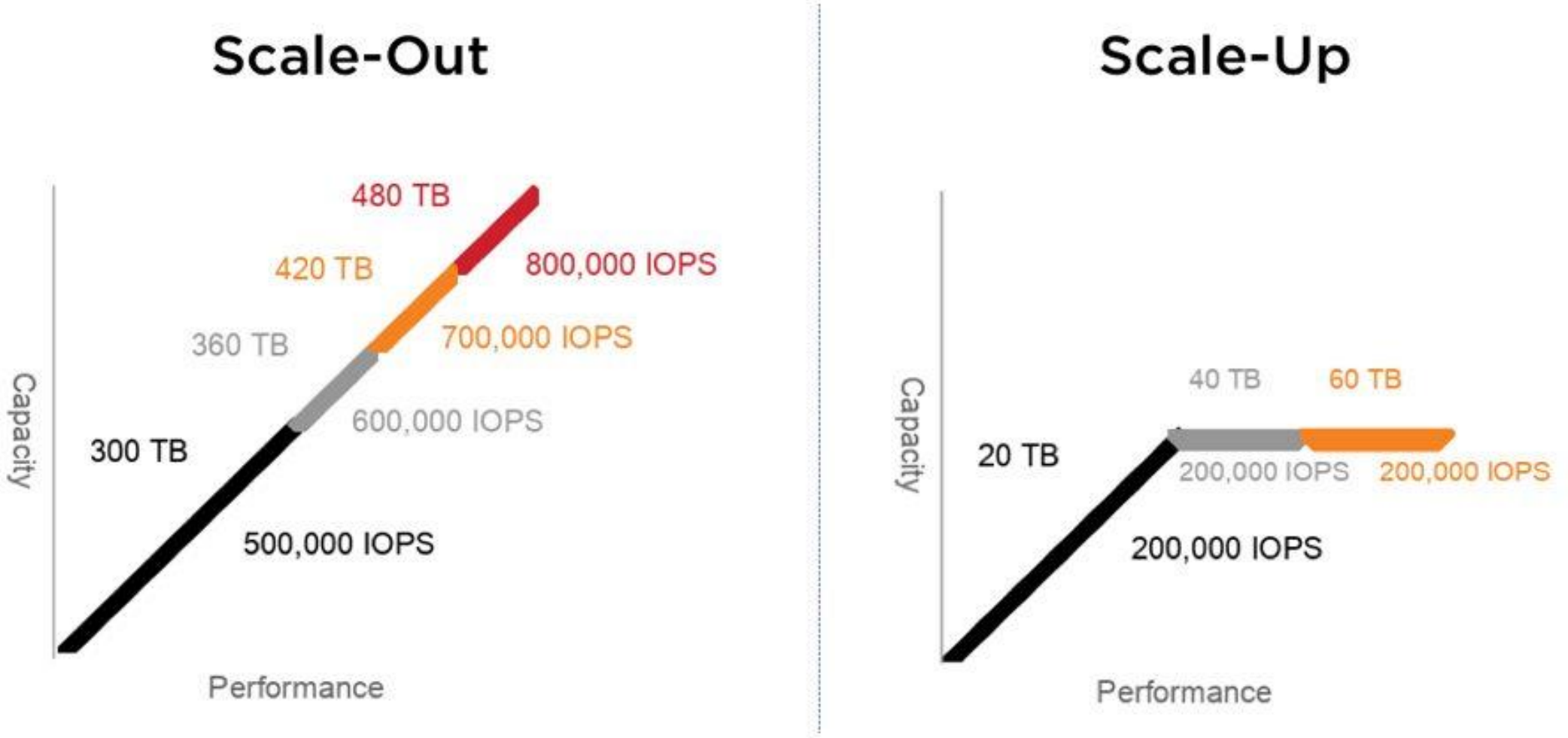


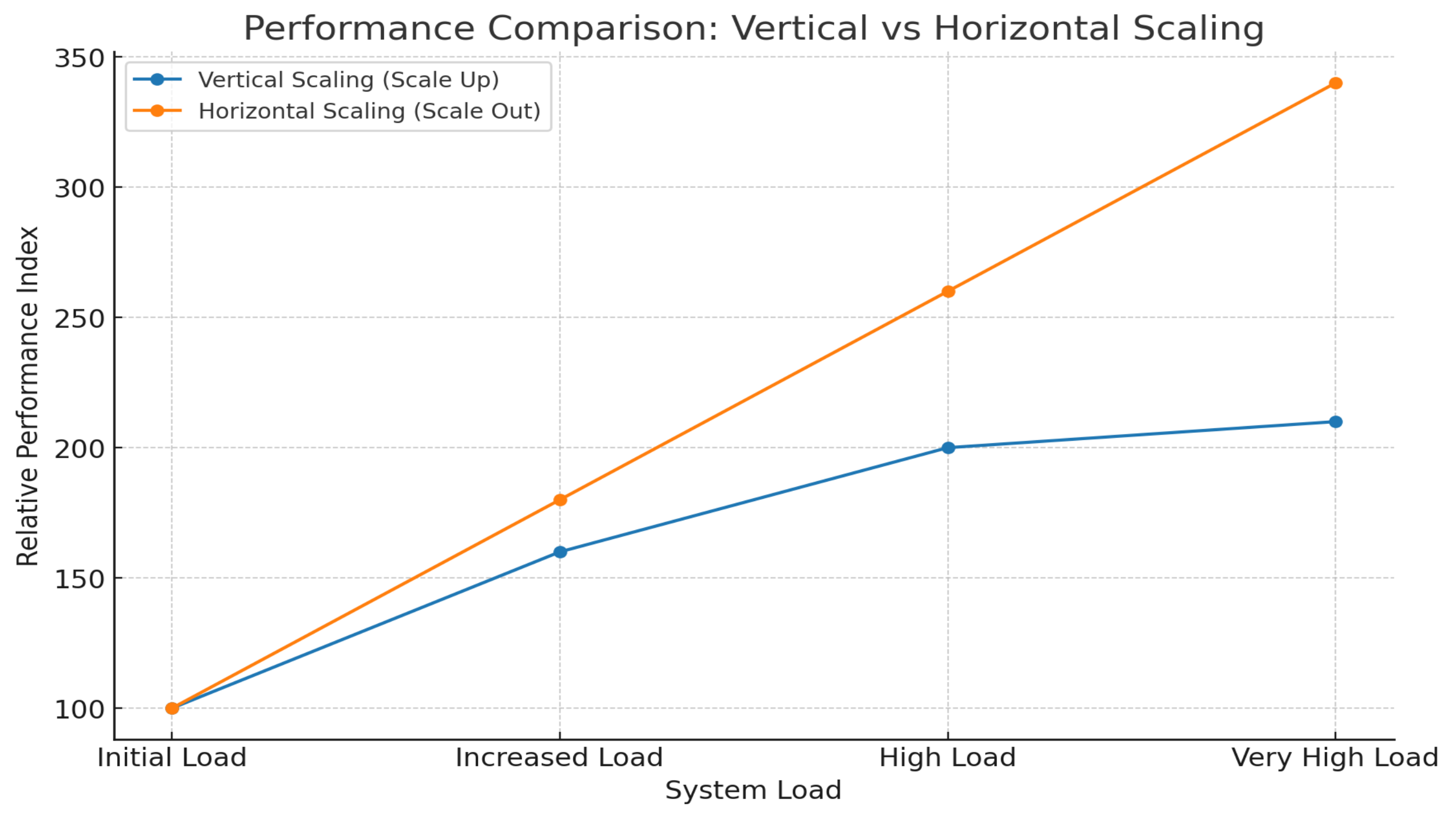
Horizontal Scaling

(Add more instances)



Scalability – Scale Up and Scale Out



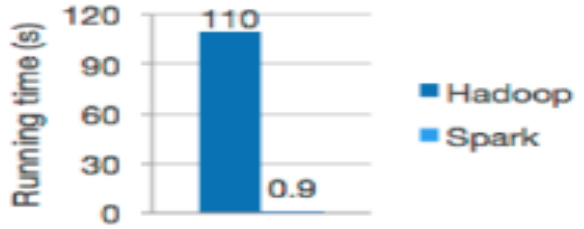

Performance Comparison: Vertical vs Horizontal Scaling
Vertical Scaling (Scale Up)
Horizontal Scaling (Scale Out)
Relative Performance Index
350
300
250
200
150
100
Initial Load
Increased Load
High Load
Very High Load
System Load

In-Memory Spark

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Spark has an advanced DAG execution engine that supports cyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala or Python.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala and Python shells.

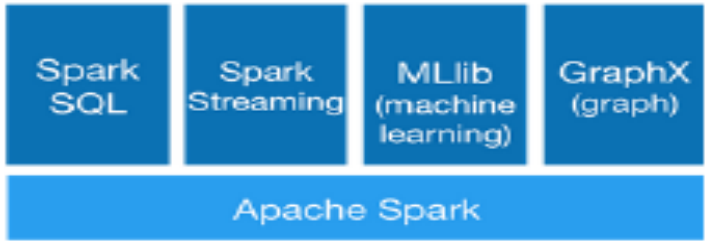
```
file = spark.textFile("hdfs://...")

file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Generality

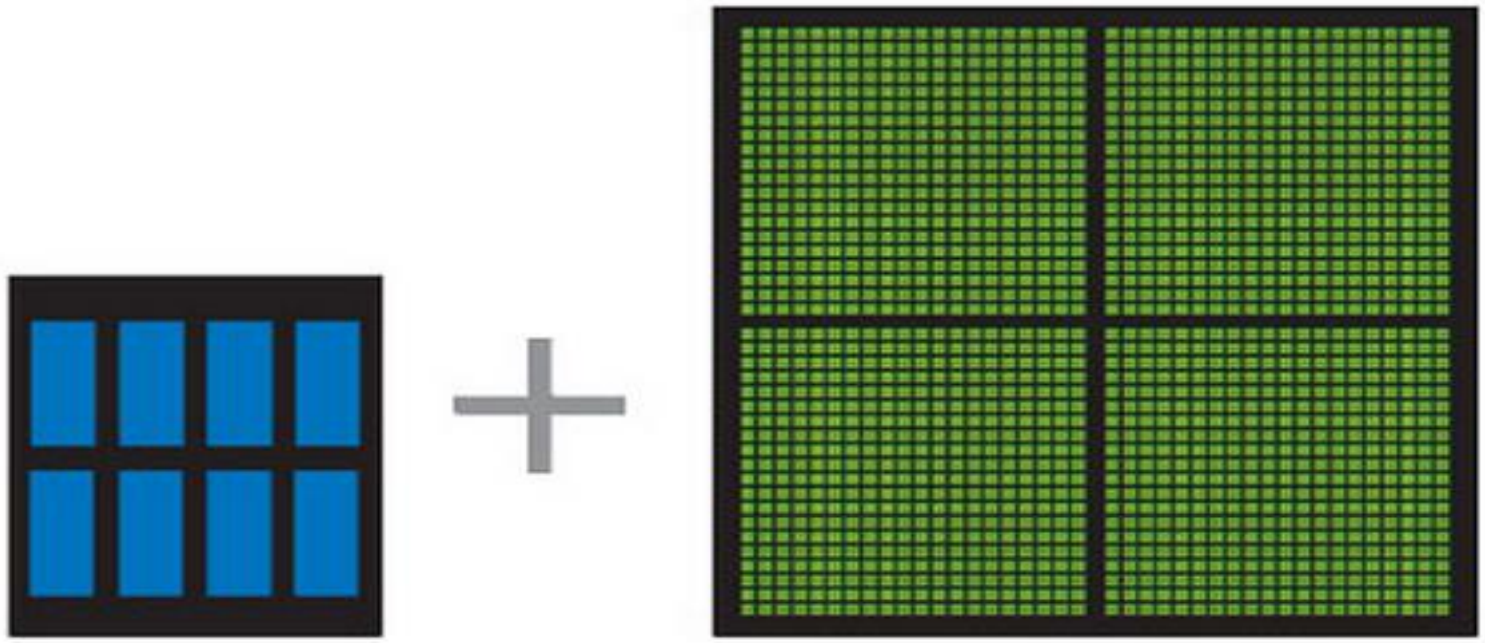
Combine SQL, streaming, and complex analytics.

Spark powers a stack of high-level tools including Spark SQL, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these frameworks seamlessly in the same application.

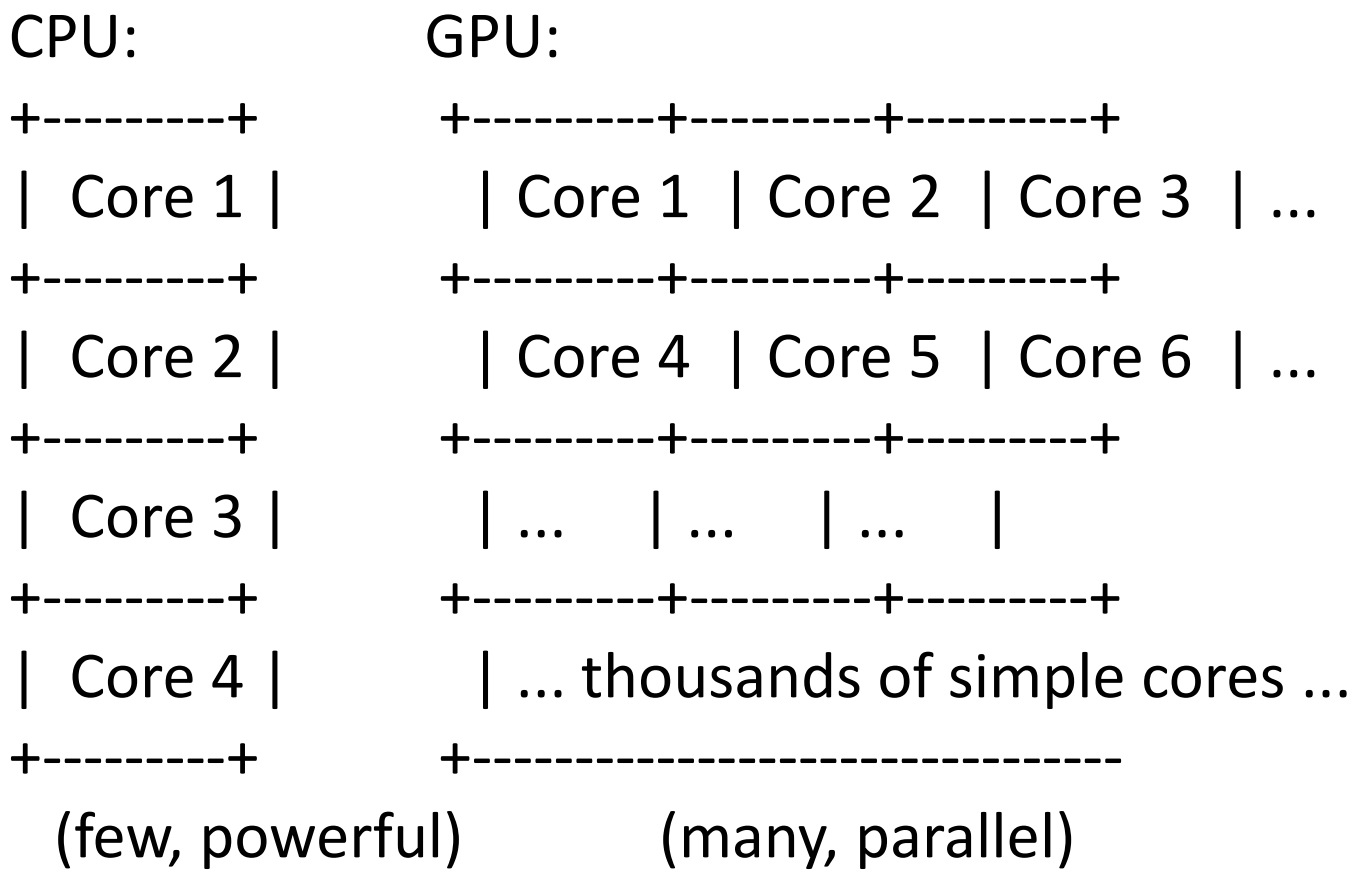


GPUs vs CPUs

GPUs have thousands of cores to process parallel workloads efficiently



GPUs vs CPUs

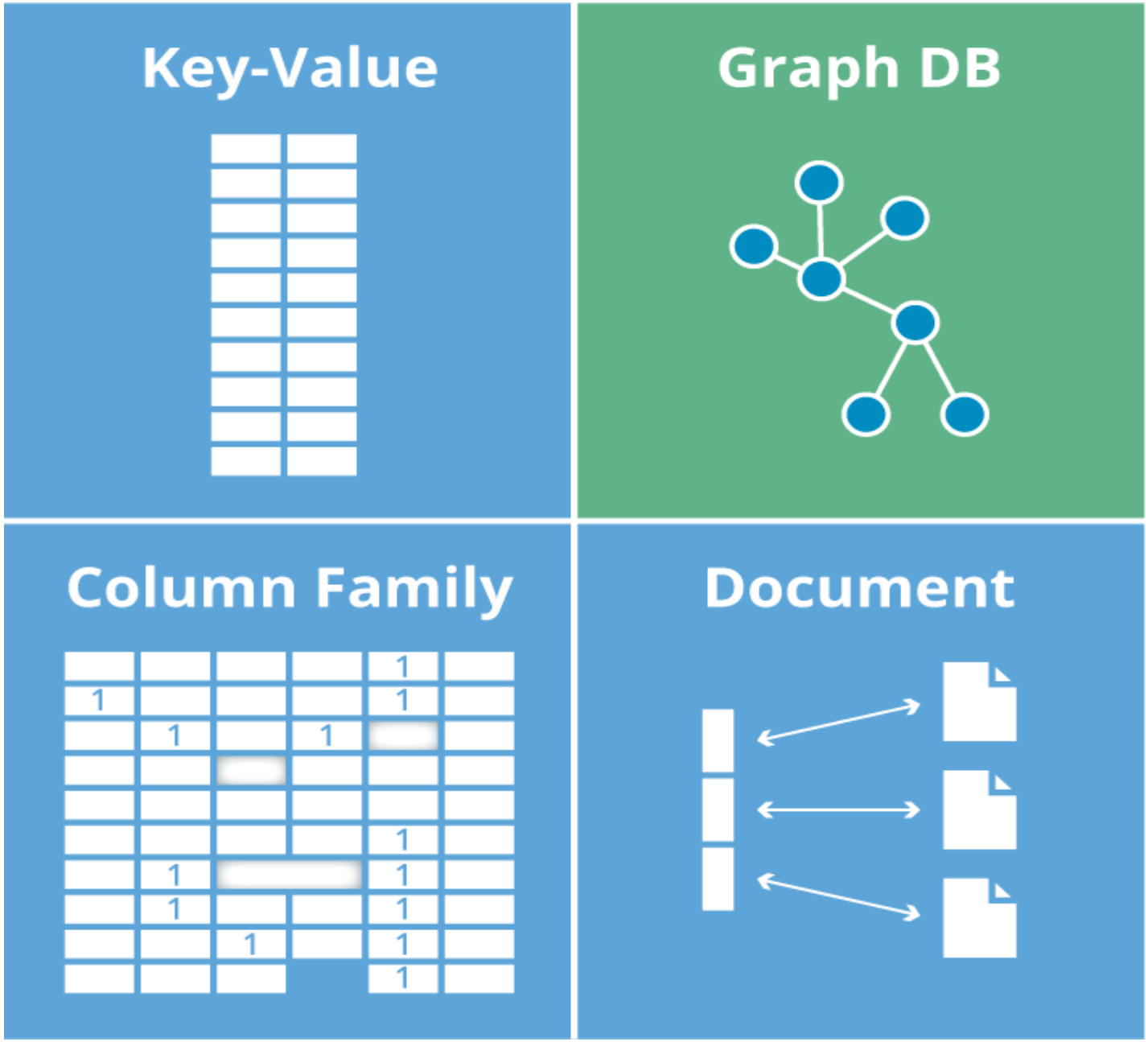

CPU:
GPU:
Core 1
Core 2
Core 3
Core 4
Core 1
Core 2
Core 3
Core 4
Core 5
Core 6
... thousands of simple cores ...
(few, powerful)
(many, parallel)

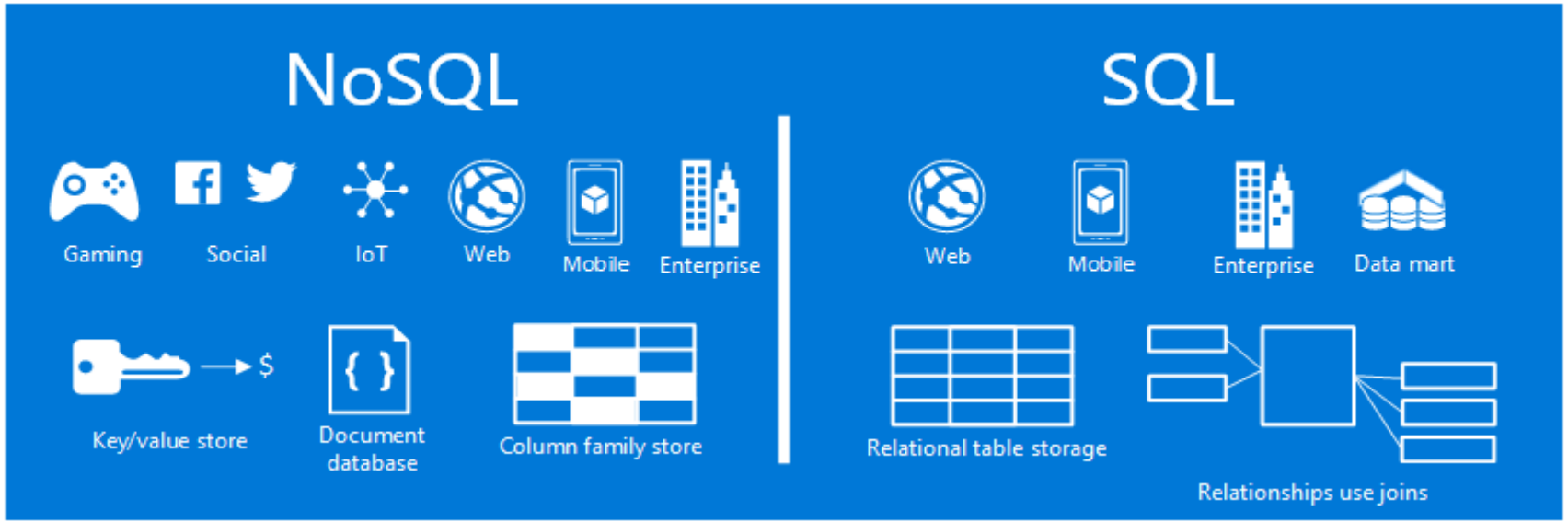
GPUs vs CPUs

Feature	**CPU** (Central Processing Unit)	**GPU** (Graphics Processing Unit)
Core Count	Few (4–32 cores in most systems)	Hundreds to thousands of cores
Core Type	Powerful, complex, general-purpose	Smaller, simpler, specialized for parallel
Best At	Sequential tasks, logic-heavy computation	Parallel tasks, matrix/vector operations
Throughput	Lower total throughput, high per-core power	Very high throughput, low per-core power
Latency	Low latency per task	Higher latency per task
Workload Examples	OS tasks, databases, sequential apps	AI/ML training, graphics, simulations
Memory	Large cache, fast access	High bandwidth VRAM (GDDR/HBM)
Flexibility	Runs a wide variety of programs	Optimized for parallel workloads

NoSQL Databases

- Stands for **N**ot **O**nly **SQL**
- Class of non-relational data storage systems
- Usually do not require a fixed table schema nor do they use the concept of joins
- All NoSQL offerings relax one or more of the ACID properties
- CAP theorem
- Document, Key-Value Pairs, Columnar Stores, Graphs
- NewSQL databases (NoSQL+ACID)


Key-Value
Graph DB
Column Family
Document


NoSQL
Gaming
Social
IoT
Web
Mobile
Enterprise
Key/value store
Document database
Column family store
SQL
Web
Mobile
Enterprise
Data mart
Relational table storage
Relationships use joins


Redis
Neo4j
the graph database
Cassandra
CouchDB
relax
mongoDB
membase
riak

Big Data Problems

- Where to store Big Data?
- When to store Big Data?
- How to store Big Data?
- How to process Big Data?
- How to analyze Big Data?
- Organization, Searching and Meta-Data
- How to manage access?
- How to copy, move and backup?
- Provenance – meta data

BIG DATA USAGE BY TECHNOLOGY

Big Data as a Service	Traditional Big Data	Traditional Database
Scalability on demand through a combination of cloud computing and distributed architecture	Scalability in processing and storage achieved through distributed architecture	Lack of resources such as computational power and storage capacity.
Virtualized data storage on a distributed platform.	Data storage on HDFS or distributed platform	Integrated hard data storage such as NAS, SAN, and traditional disks
Structured and unstructured data on cloud environment	Structured and unstructured data	Structured data
Advanced analytics functions with on-demand computing power	Advanced analytics functions	Reporting using tools such as OLAP
Ubiquitous accessibility	Limited accessibility	Limited accessibility
Analytical capability derived from out-of-box domain-specific algorithms along with custom coding	Analytical capability derived through custom coding	Analytical capability derived through custom coding

DIFFERENCE BETWEEN BIG DATA AND INTERNET OF THINGS

IoT Analytics 3-Tier Architecture

using the example of a wind turbine farm

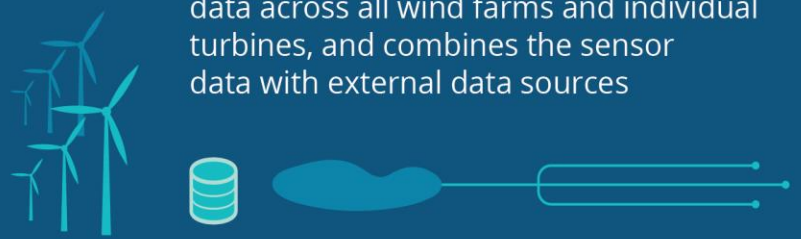
TIER 1 Performs individual wind turbine real-time performance analysis and optimization

100 99 99 99
100 99 100 99
99 100 81 82
100 99 99 98
100 99 99 99

Purpose-built T1 edge gateways leverage real-time data compression techniques to only send a subset of the critical data

TIER 2 Optimizes performance and predicts maintenance needs across the wind turbines in the same wind farm

TIER 3 The data-lake-enabled core analytics platform that aggregates the critical data across all wind farms and individual turbines, and combines the sensor data with external data sources



To drive meaningful business impact, you will need to begin with the business and not the technology:

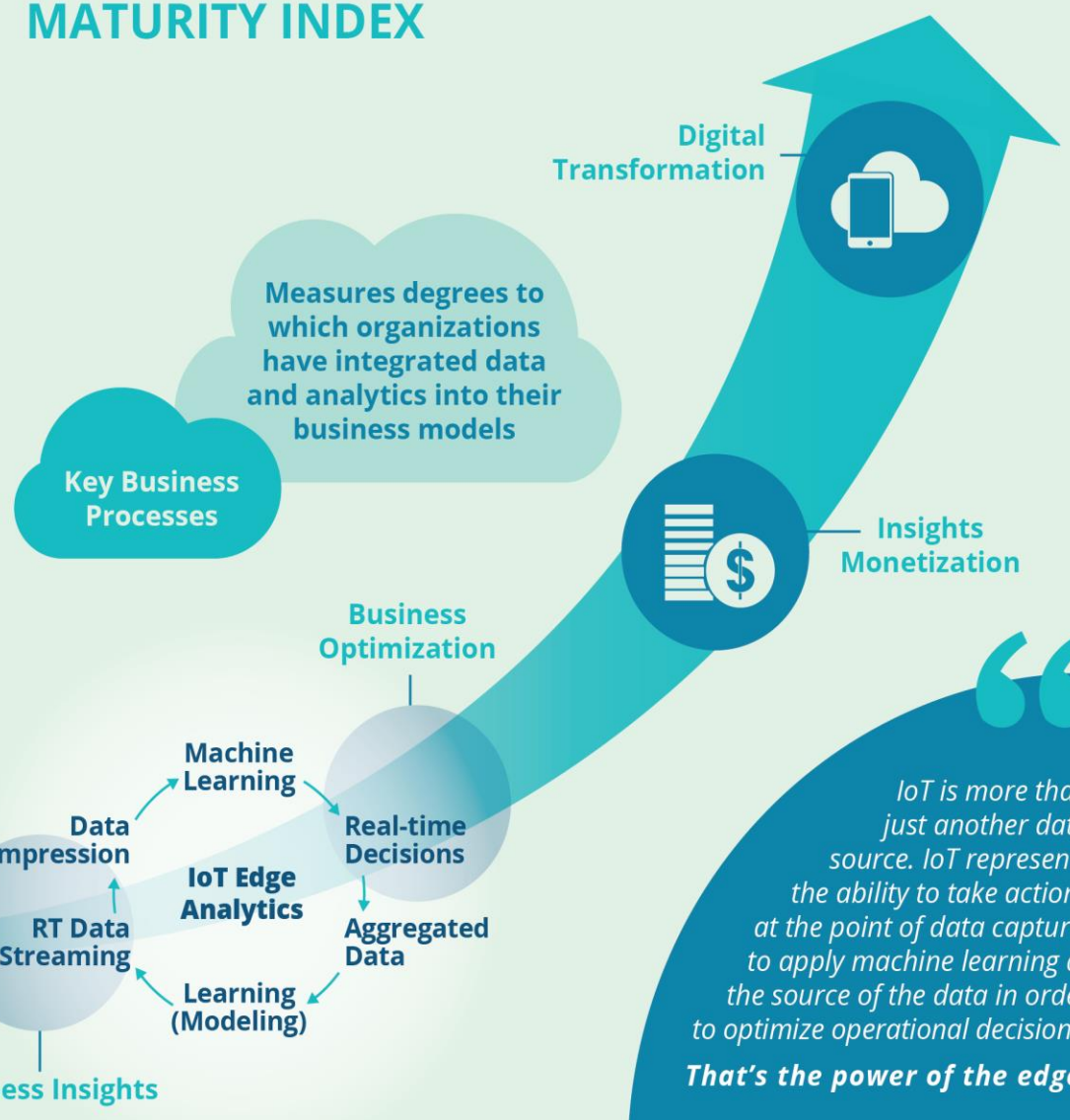
Engage the business stakeholders on day one

Align the business and IT teams

Understand the organization's key business and operational initiatives

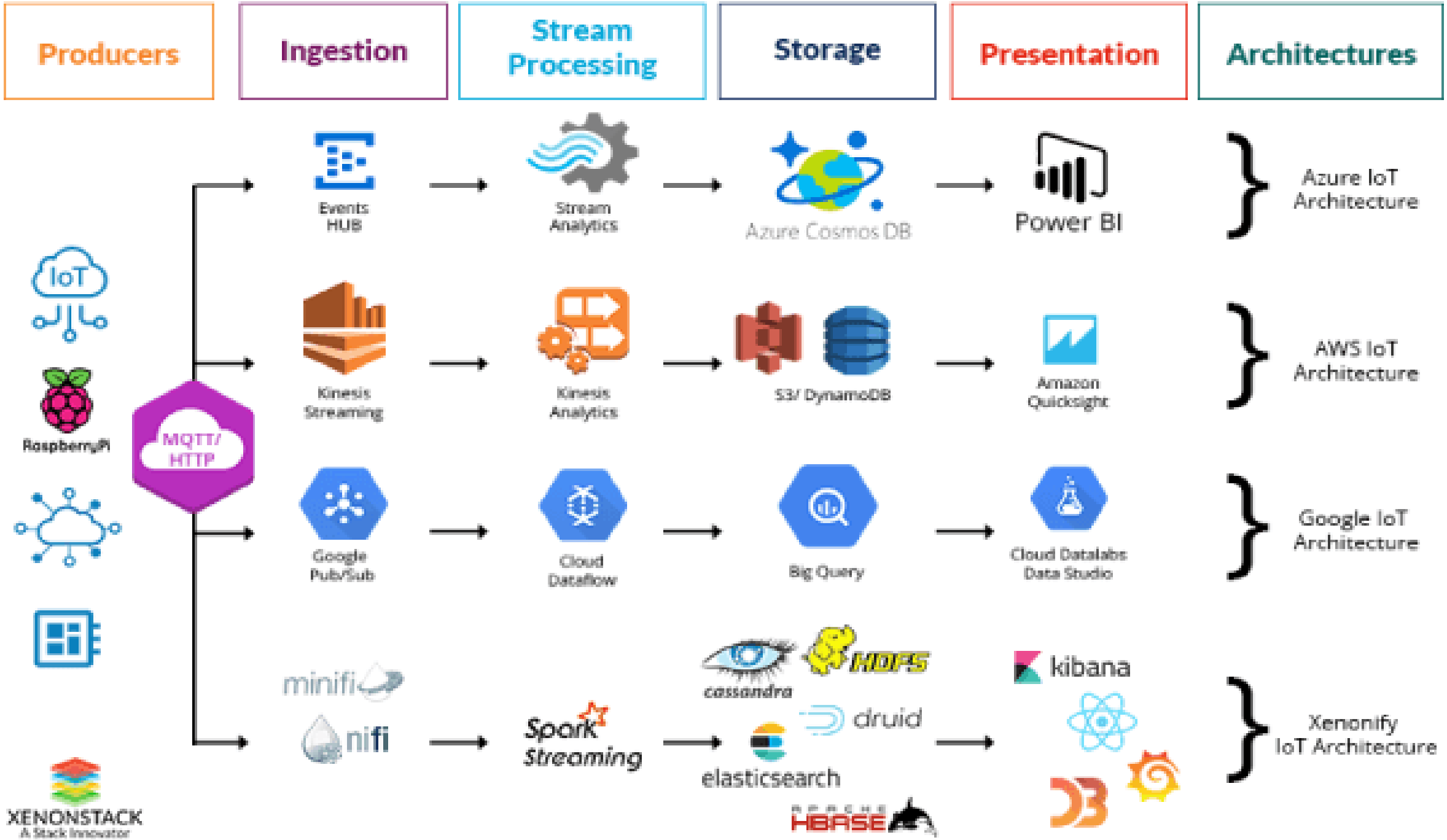
Identify and prioritize the use cases (decisions/goals) that support those business initiatives

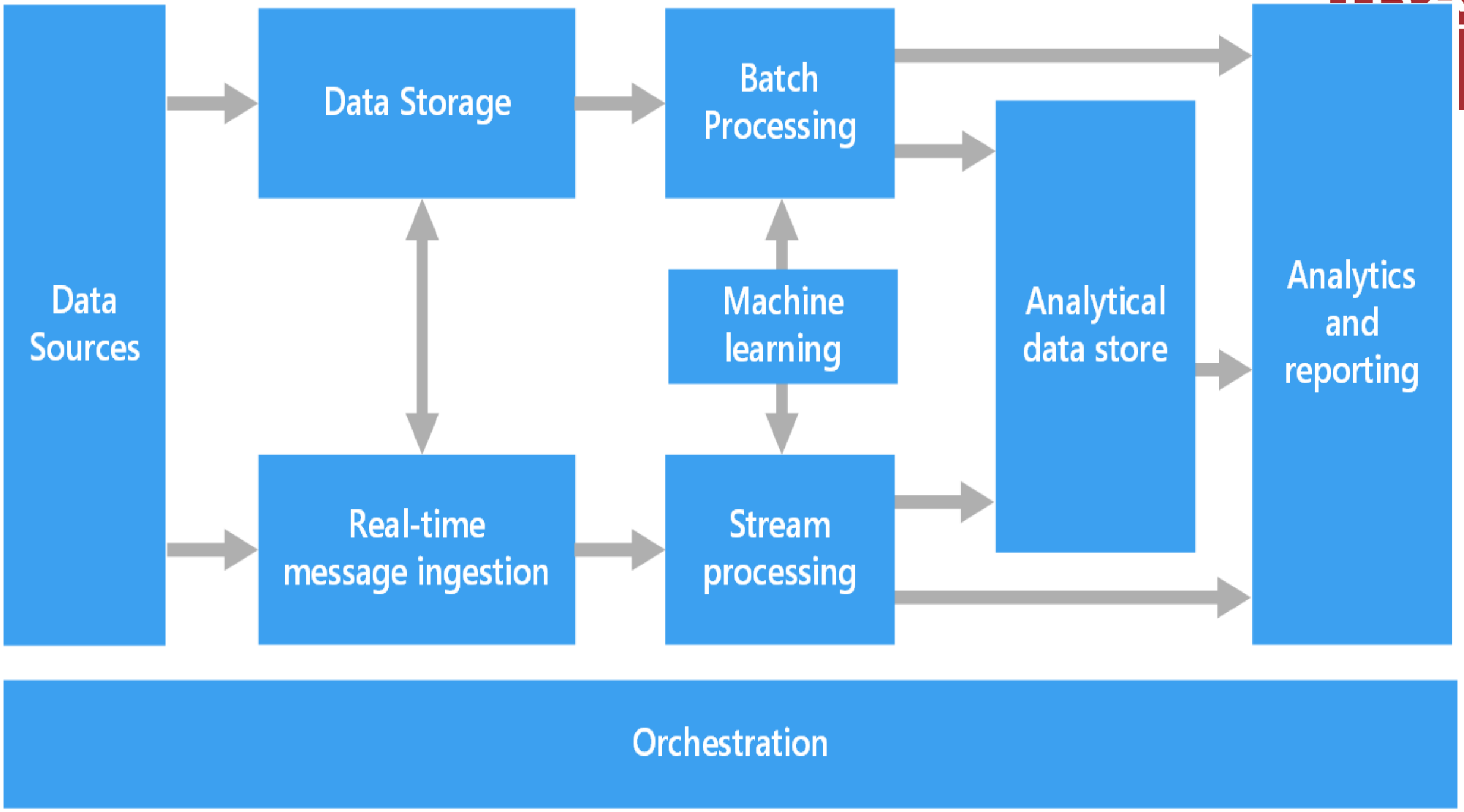
BIG DATA BUSINESS MODEL MATURITY INDEX



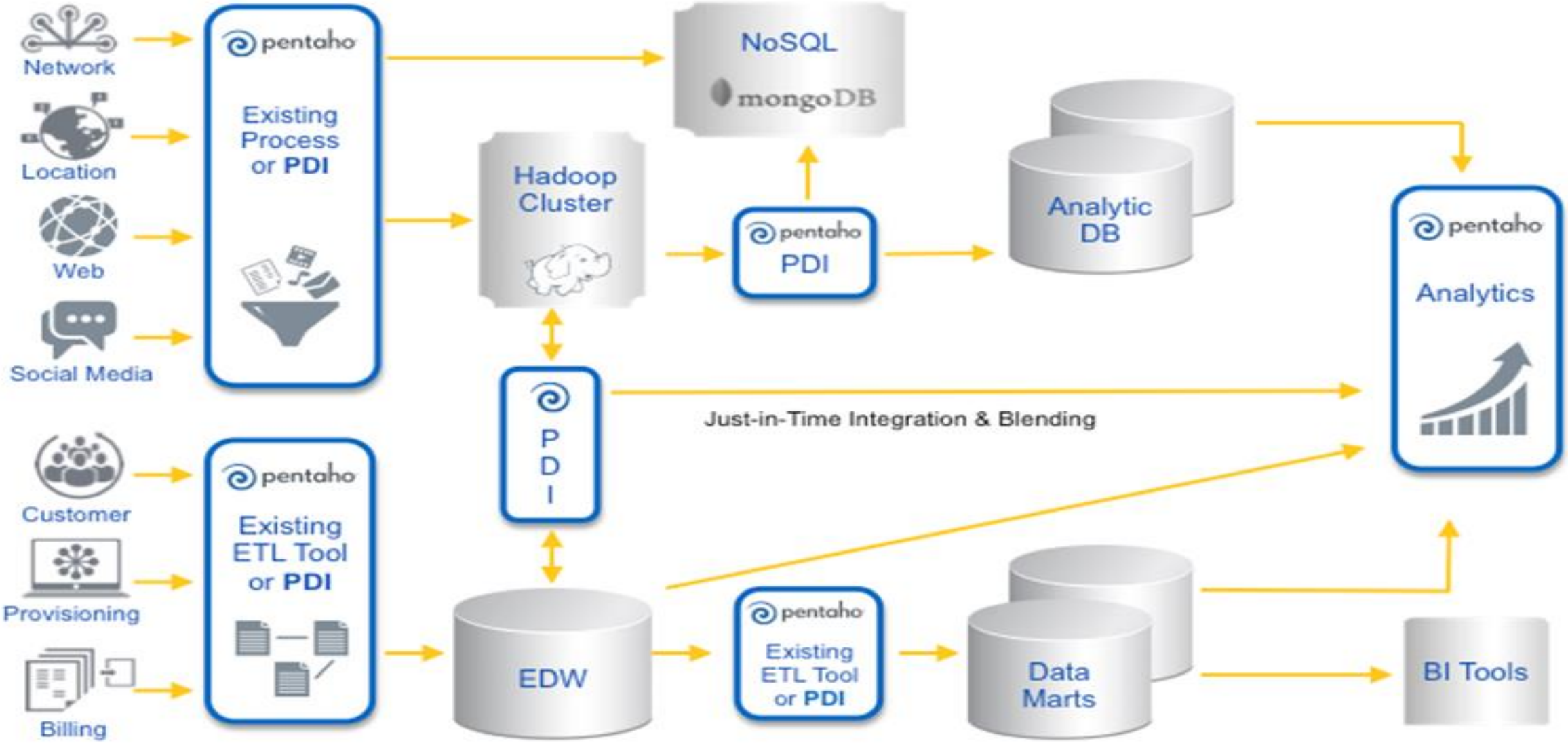
IoT is more than just another data source. IoT represents the ability to take actions at the point of data capture; to apply machine learning at the source of the data in order to optimize operational decisions.

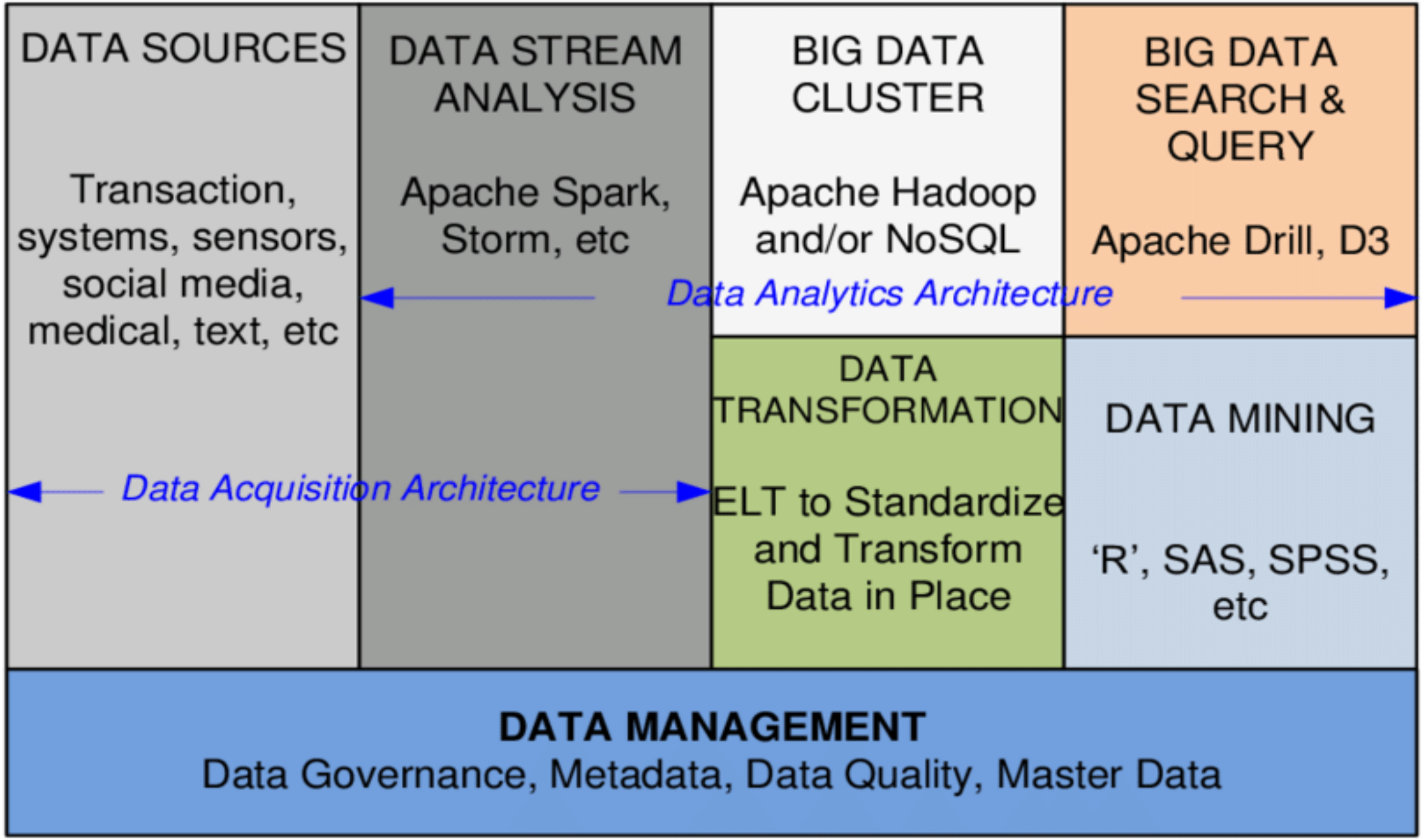
That's the power of the edge.


Producers
Ingestion
Stream Processing
Storage
Presentation
Architectures
IoT
RaspberryPi
MQTT/ HTTP
Events HUB
Stream Analytics
Azure Cosmos DB
Power BI
Azure IoT Architecture
Kinesis Streaming
Kinesis Analytics
S3/ DynamoDB
Amazon Quicksight
AWS IoT Architecture
Google Pub/Sub
Cloud Dataflow
Big Query
Cloud Datalabs Data Studio
Google IoT Architecture
minifi
nifi
Spark Streaming
cassandra
HDFS
druid
elasticsearch
HBASE
kibana
Xenonify IoT Architecture
XENONSTACK
A Stack Innovator


Data Sources
Data Storage
Batch Processing
Machine learning
Analytical data store
Analytics and reporting
Real-time message ingestion
Stream processing
Orchestration

Evolving Big Data Architectures

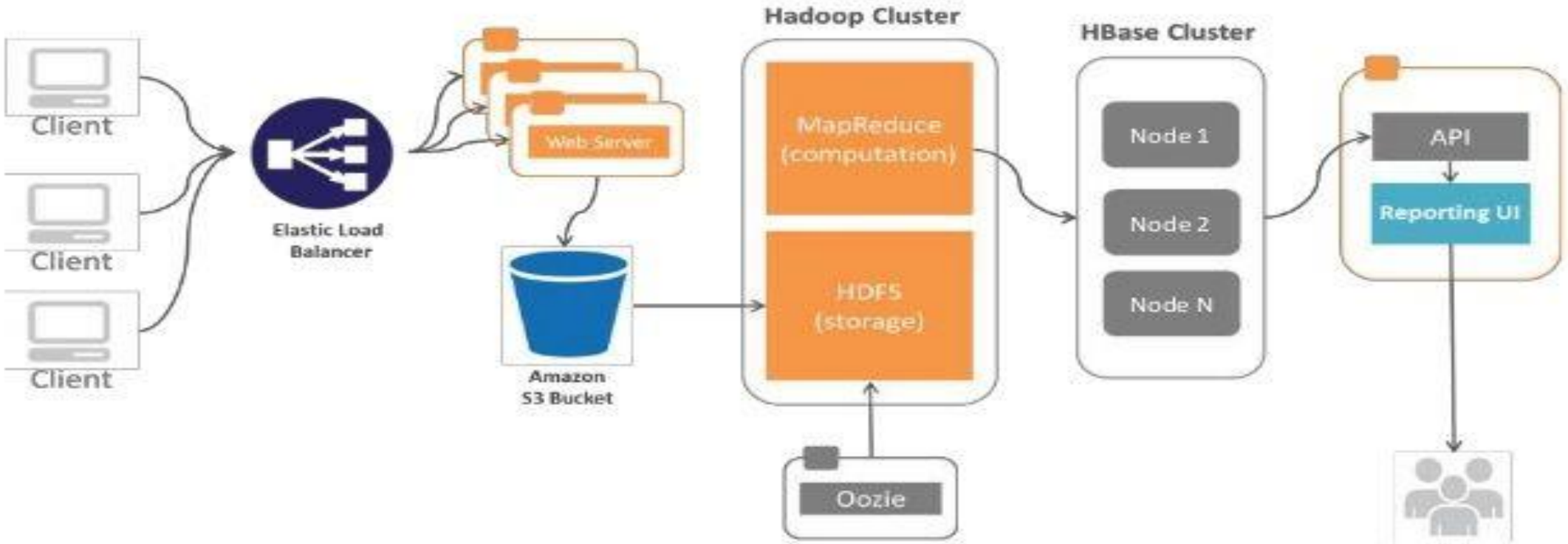



DATA SOURCES
Transaction, systems, sensors, social media, medical, text, etc
DATA STREAM ANALYSIS
Apache Spark, Storm, etc
BIG DATA CLUSTER
Apache Hadoop and/or NoSQL
BIG DATA SEARCH & QUERY
Apache Drill, D3
Data Analytics Architecture
Data Acquisition Architecture
DATA TRANSFORMATION
ELT to Standardize and Transform Data in Place
DATA MINING
'R', SAS, SPSS, etc
DATA MANAGEMENT
Data Governance, Metadata, Data Quality, Master Data

Solution Architecture

Technologies:

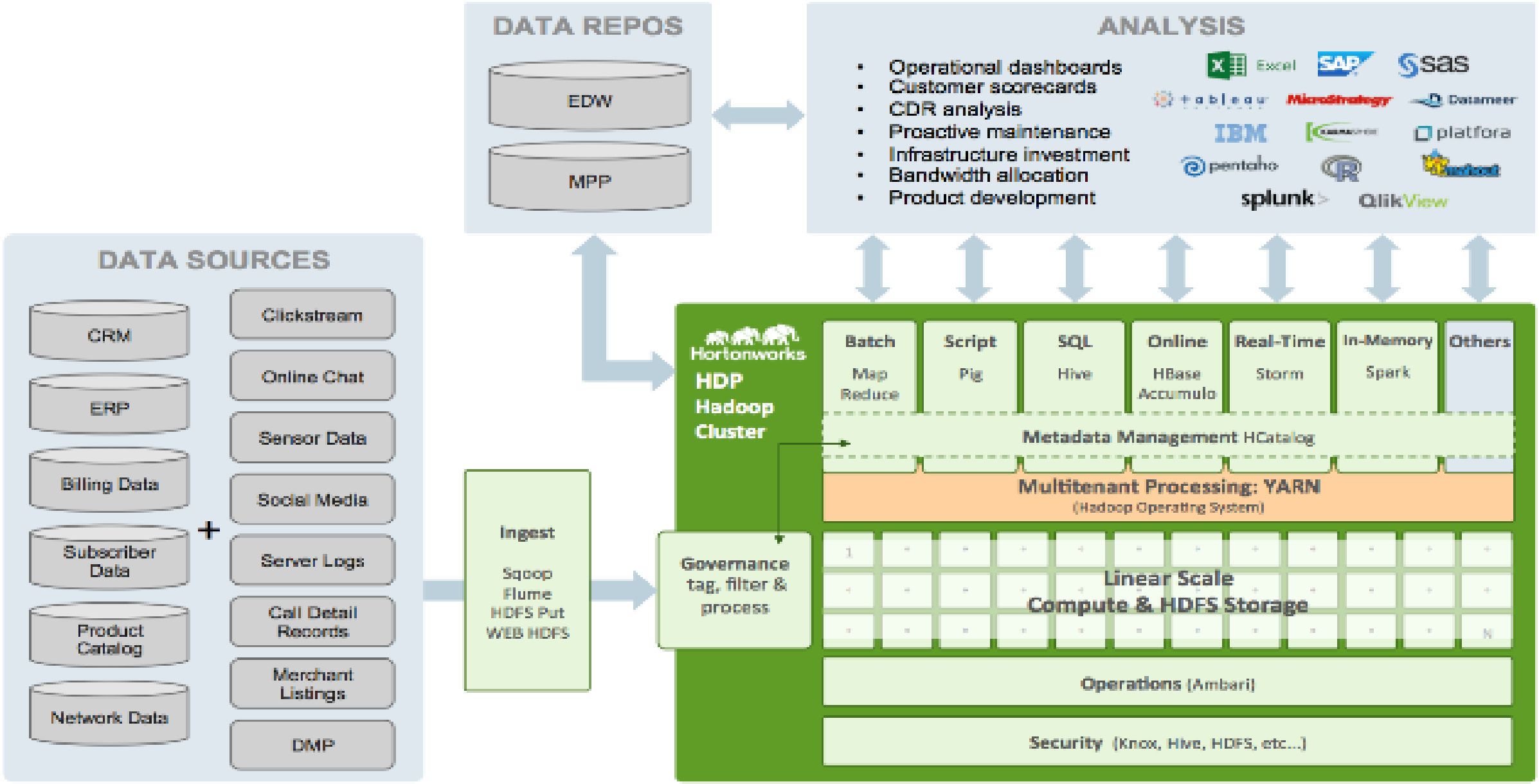
- Amazon S3
- Flume
- Hadoop/HDFS, MapReduce
- HBase
- Oozie
- Hive



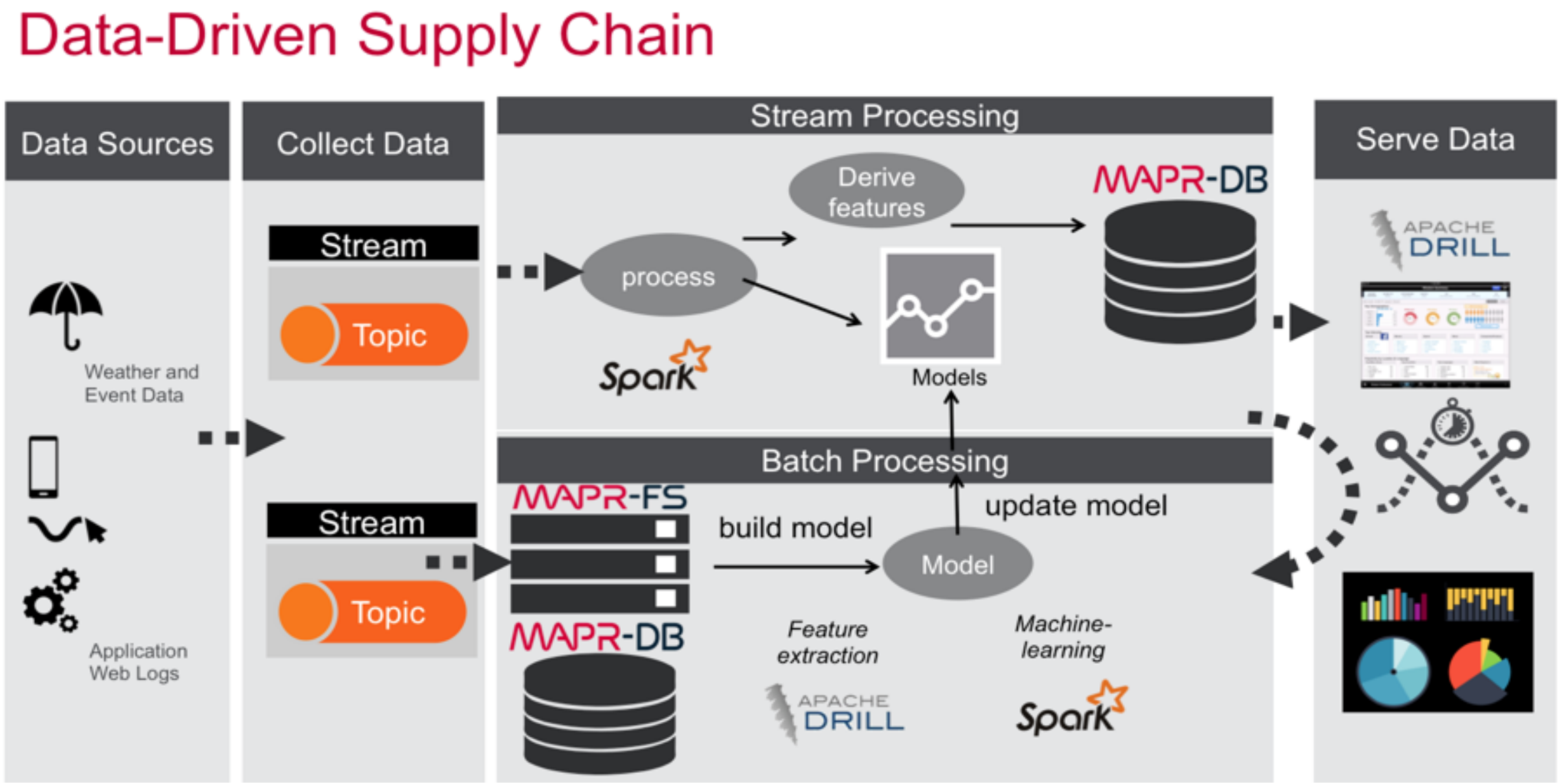
SoftServe

24

Big Data Process for Telecom

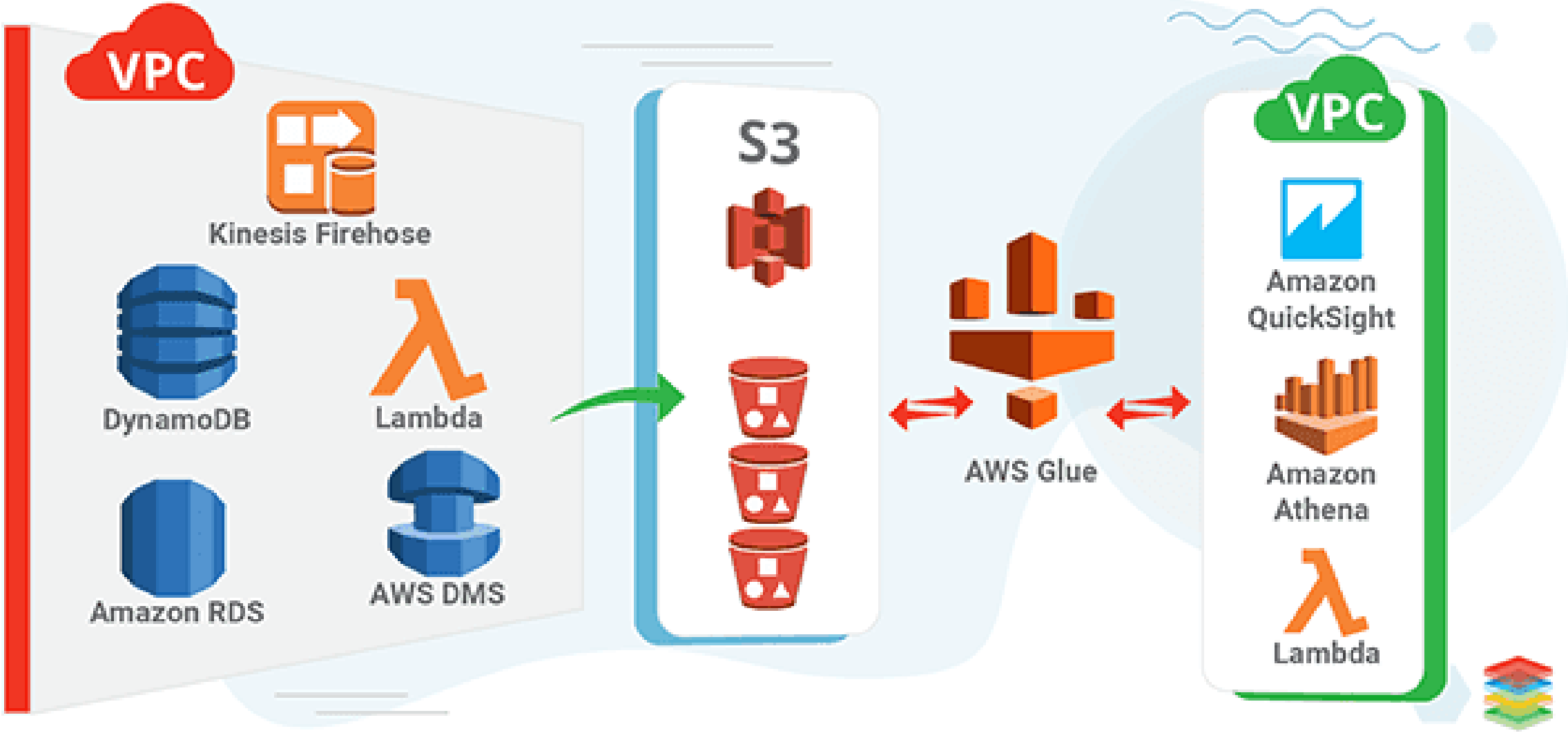


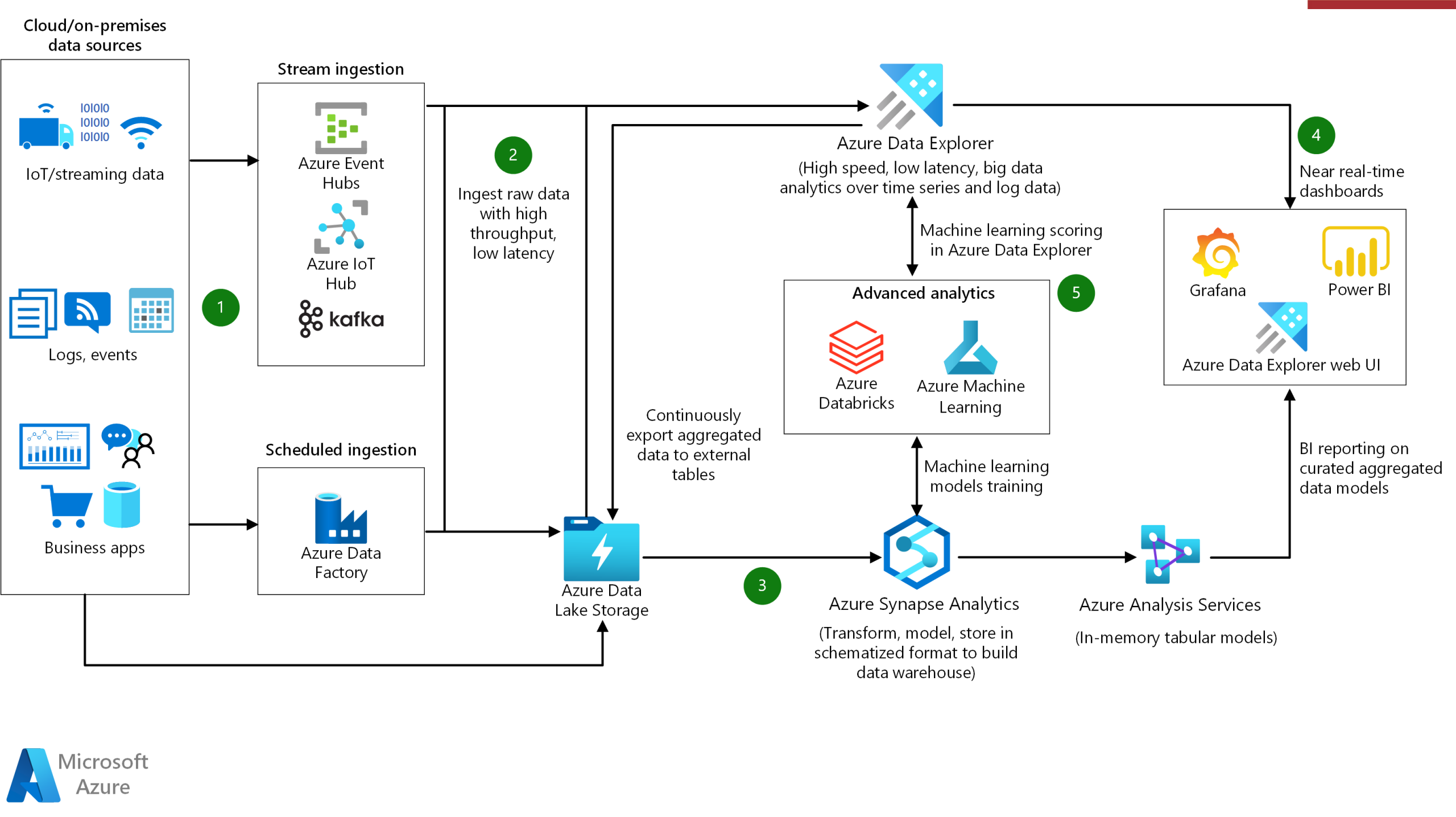
Big Data Process for Retail



Location Tracking

Big Data Pipeline on AWS




Cloud/on-premises data sources
IoT/streaming data
Logs, events
Business apps
1
Stream ingestion
Azure Event Hubs
Azure IoT Hub
kafka
Scheduled ingestion
Azure Data Factory
2
Ingest raw data with high throughput, low latency
Azure Data Explorer
(High speed, low latency, big data analytics over time series and log data)
Machine learning scoring in Azure Data Explorer
Advanced analytics
Azure Databricks
Azure Machine Learning
5
Machine learning models training
Continuously export aggregated data to external tables
Azure Data Lake Storage
3
Azure Synapse Analytics
(Transform, model, store in schematized format to build data warehouse)
Azure Analysis Services
(In-memory tabular models)
4
Near real-time dashboards
Grafana
Power BI
Azure Data Explorer web UI
BI reporting on curated aggregated data models
Microsoft Azure