Big Data Analytics

Module-1
Introduction to Big Data

Prof. Lokeshkumar R


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY
New Craigslist Ads
320+ NEW twitter ACCOUNTS
13,000+ iPhone APPLICATIONS DOWNLOADED
100+ NEW Linked in ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
600+ NEW VIDEOS
50+ WordPress DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
60+ NEW BLOGS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
1,500+ BLOG POSTS
79,364 WALL POSTS
510,040 COMMENTS
Google
Google Search
GO-Globe.com

Data never sleeps…

How Much Data Is Generated Every Minute? 24/7/365

GPT-3 – **3.1 million** / per minute

GPT-4.0 - **6 million** / per minute

50 M to 100 M words generated per minute globally

Email Users Send **204,166,667** Emails

Google Receives Over **2,000,000** Search Queries

Apple Receives About **47,000** App Downloads

Brands on Facebook Get **34,722** Likes

Big Data Facts

According to McKinsey – a reta
iler using big data to the fullest
could increase its operating
margins by more than **60**

According to Zuckerberg, **8 bill
ion** pieces of content are share
d via Facebook's Open Graph

Bad data or poor quality data
costs US businesses **$900
Billion** annually

Google's Sundar claims that
every **two days** now we
create as much informati
s we did from the dawn o
lization until **2003**

According to Gartner Big Data
will drive **$900 Billion** in
ending through 2025

By **2025 9.4 Million** IT jobs
d to support Big D
ting **1.9 million**
nited States

What is Big Data?

Definition

Big data refers to the large volume of structured, semi-structured, and unstructured data that inundates businesses on a day-to-day basis.

Sources

It is generated from various sources such as social media, business transactions, sensors, and other digital sources.

Importance

Understanding big data is crucial for organizations to uncover hidden patterns, correlations, and other insights.

Data → Big & Fast Data

- The data **volume** gets bigger (in the range of petabytes, exabytes Zettabyte (ZB), Yottabyte (YB), Xenottabyte
- The data generation, transmission and crunching frequency have gone up significantly
- The data structure too has got diversified (poly-structured data, the variety)
- The correctness, / accuracy, timeliness, and trustworthiness of data (veracity)
- The intriguing and interesting relationships between data items (viscosity)

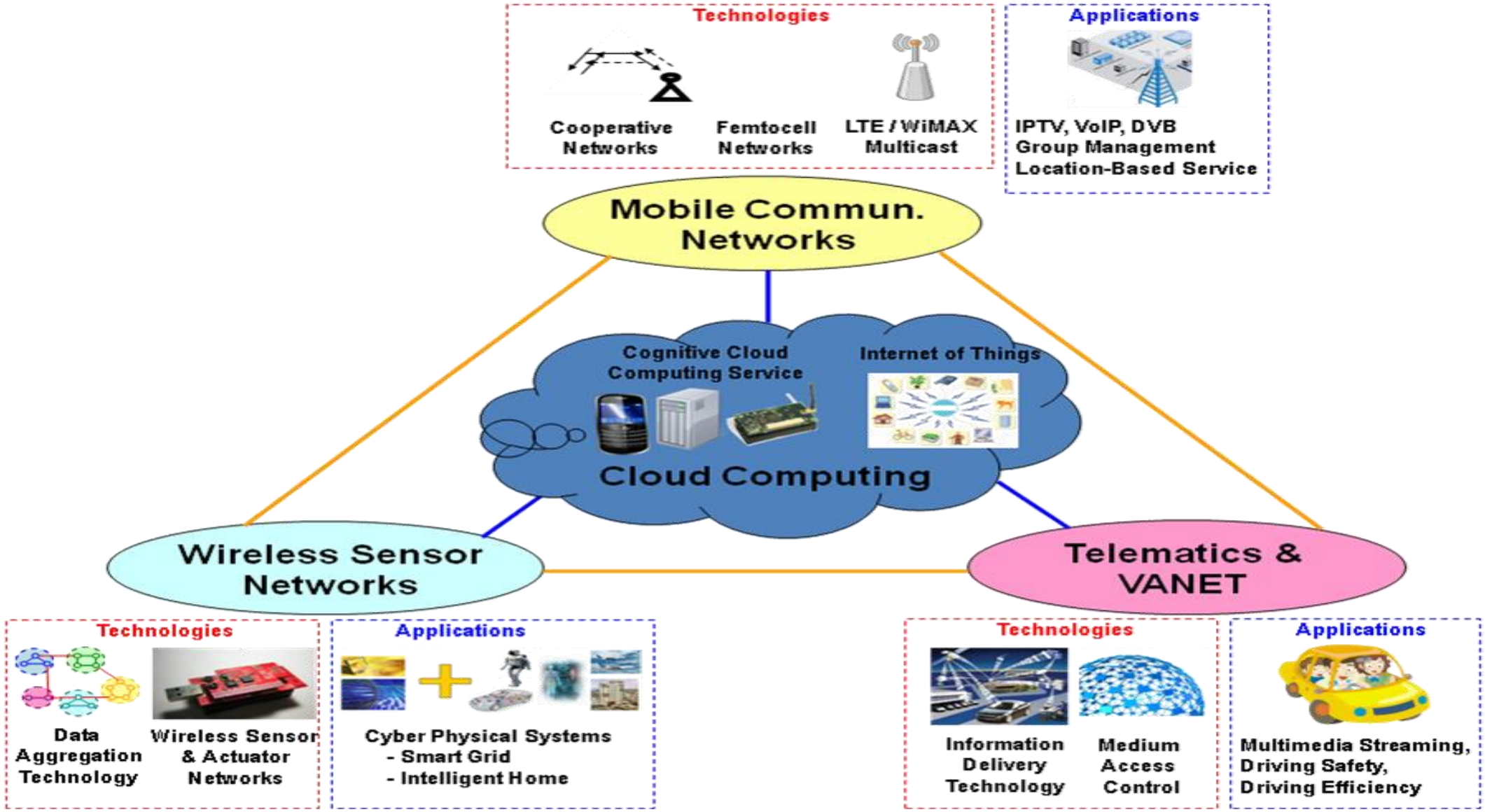
The Key Drivers for Big & Fast Data

Primarily due to newer Data Sources, Devices & Emerging Technologies

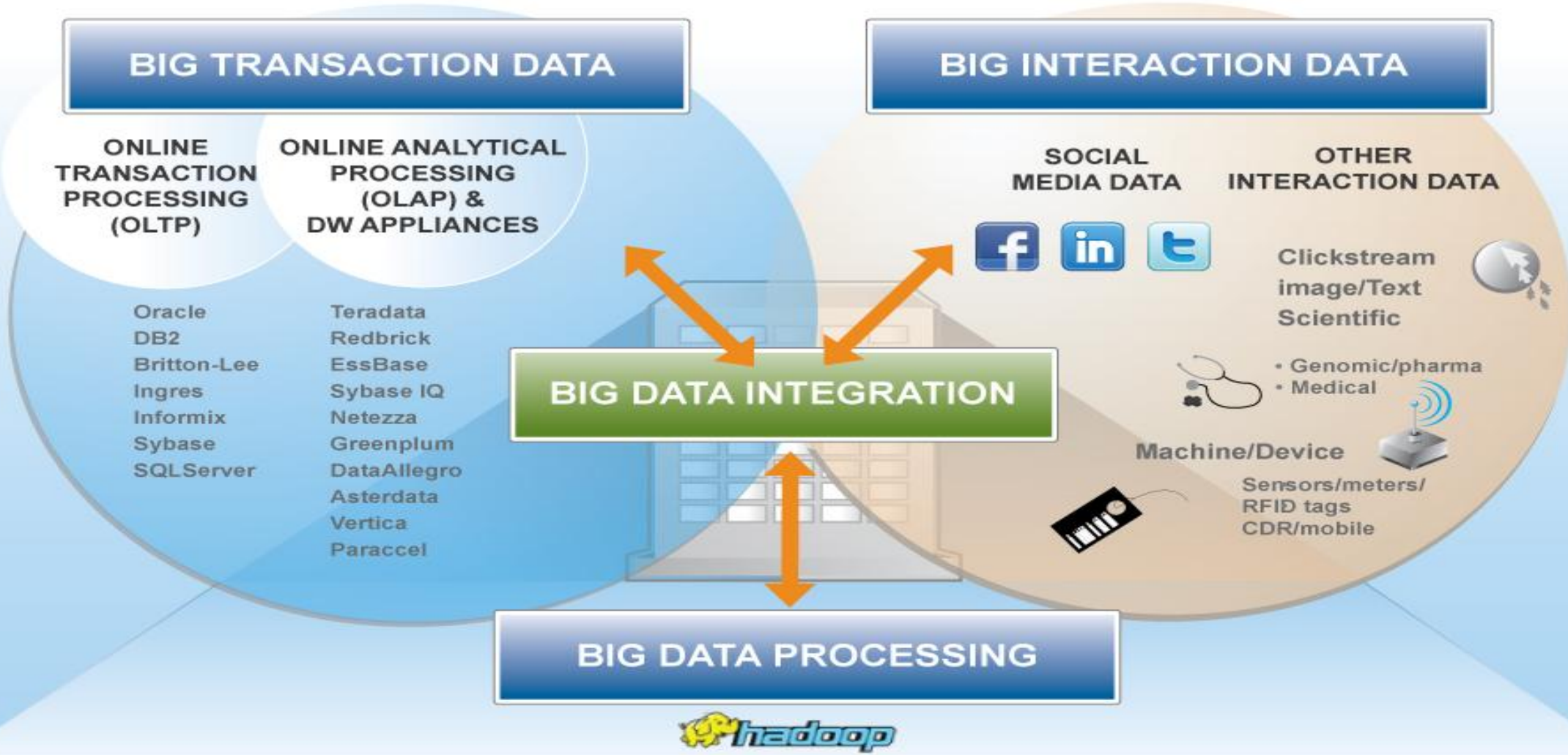
- Sentient Materials / Smart Objects / Digitized Entities through deeper digitization enabled by edge technologies (Nano and micro-scale sensors, actuators, codes, chips, controllers, specks, smart dust, tags, stickers, LED, etc.)
- Connected, resource-constrained, and embedded devices and machines (Device integration buses, data transmission protocols, operating systems, etc.)
- Ambient Sensing, Vision, and Perception Technologies
- Social media and Professional and Knowledge Sharing Sites
- Consumerization (Mobiles and Wearables)
- Centralization, Commoditization, Containerization & Industrialization (Cloud Computing)
- Communication (Ambient, Autonomic and Unified)
- Integration (D2D, D2E, D2C, C2C, etc.)
- Big & Fast Data Platforms and Infrastructures

The convergence of technologies lays a profound foundation for Large-scale Data Generation

The extreme connectivity enables data generation in heaps



Definition: Big data is the confluence of the three trends consisting of Big Transaction Data, Big Interaction Data and Big Data Processing



The Next-Gen Analytical Capabilities

The Emergence of Newer Analytics and Applications (As per McKinsey, at least 13 industry verticals got identified to gain immense benefits out of big data analytics)

Generic (Horizontal)	Specific (Vertical)
Real-time Analytics	Social Media Analytics
Predictive Analytics	Operational Analytics
Prescriptive Analytics	Machine Analytics
High-Performance Analytics	Retail and Security Analytics
Diagnostic Analytics	Sentiment Analytics
Descriptive Analytics	Security & Fraud Analytics
Personalized Analytics	Weather Analytics
Stream Analytics	Watson Content Analytics

Need of Big Data

Emerging Trends

Current Trends in Big Data Analytics

Unveiling the Future of Big Data Analytics

AI Integration

AI advancements will drive predictive analytics to new heights, enhancing decision-making processes.

Machine Learning Evolution

Machine learning algorithms will become more sophisticated, enabling deeper insights

Real-time Data Processing

Real-time analytics will empower businesses to react swiftly to market changes and consumer

Edge Computing Expansion

Edge computing will expand to process data closer to the source, reducing latency and

Blockchain Interoperability

Enhanced blockchain integration will ensure secure and transparent data sharing across

SWOT Analysis

Challenges in Implementing Big Data Analytics

Strategies to Overcome Data Privacy, Integration Issues, and Skill Gaps

Implementing Robust Data Governance, Training Programs, Collaboration with Experts

To overcome these challenges, organizations can implement robust data governance frameworks, conduct targeted training programs to bridge skill gaps, and collaborate with experts in the field.

Data Privacy, Integration Issues, Skill Gaps

This section discusses the challenges related to data privacy concerns, integration issues with existing systems, and skill gaps within organizations.

Technological Advancements

Future of Big Data Analytics

Unveiling the Transformative Power Ahead

2022

Increased adoption of AI-driven analytics tools.

2023

Rapid growth in the utilization of IoT data for analytics purposes.

2024

Mainstream integration of machine learning algorithms in big data analysis.

2024

Emergence of real-time analytics as a standard practice.

2025?

AI-Driven Analytics	Integration of AI/ML with Big Data for real-time, predictive, and prescriptive insights.
Edge + Cloud Synergy	More data processing at the edge (IoT, devices) with cloud platforms for storage & analytics.
Data Democratization	Self-service analytics tools enabling non-technical users to explore data.
DataOps & Automation	Automated data pipelines and governance using tools like Apache Airflow, dbt, and MLflow.
Privacy-First Analytics	Adoption of privacy-enhancing technologies like federated learning, synthetic data, differential privacy.
Synthetic Data Growth	Used for AI training, simulation, and testing when real data is limited or sensitive.

Efficiency & Scalability vs. Insights

Comparing Traditional and Big Data Analytics

Enhancing Insights and Efficiency through Modern Data Analytics

Efficiency

- Traditional analytics: Limited by data volume and processing speed.
- Big Data Analytics: Processes vast amounts of data rapidly for quick insights.

Scalability

- Traditional analytics: Limited scalability due to infrastructure constraints.
- Big Data Analytics: Highly scalable, capable of handling massive datasets.

Harnessing Big Data

- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

Characteristics of Big Data

The Model Has Changed…

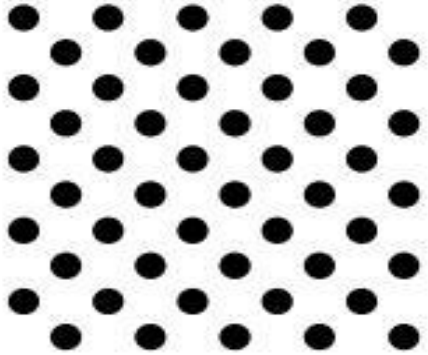
- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

Some Make it 4V's

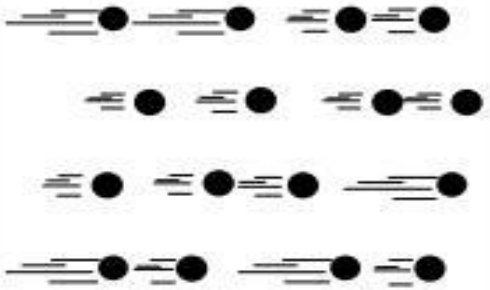
Volume



Data at Rest

Terabytes to exabytes of existing data to process

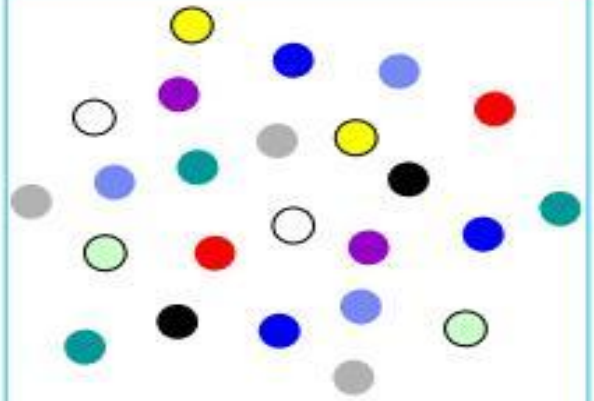
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

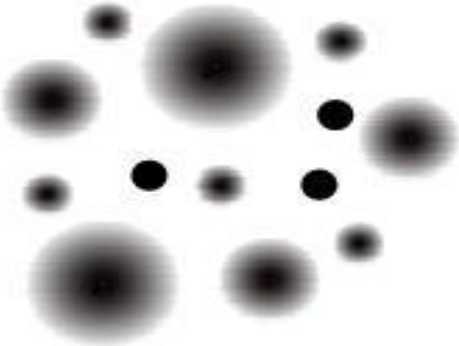
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

New Set of Characteristics in after 2022

Volume: The Scale of Data

Exploring the Implications of Data Volume

1

Large datasets from various sources

Data volume refers to the massive amount of information generated from diverse origins, including IoT devices, social media platforms, and enterprise systems.

2

Importance of storage and processing capabilities

Managing data at scale necessitates robust storage solutions and advanced processing capacities to handle the sheer volume efficiently.

3

Examples: Social media data, transaction records

Illustrative data sources such as social media content and transactional records exemplify the substantial volume of information that organizations deal with daily.

4

Large datasets from various sources

Data volume refers to the massive amount of information generated from diverse origins, including IoT devices, social media platforms, and enterprise systems.

Velocity Insights —

Velocity: The Speed of Data

Unlocking the Power of Real-Time Data Processing

- Real-time data processing
- Impact on decision-making
- Examples: Streaming data from IoT devices
- Examples: Financial market data

Variety: The Different Forms of Data

Exploring Various Types and Sources of Data

Structured Data

Data that adheres to a pre-defined data model or schema, simplifying storage and analysis processes.

Semi-Structured Data

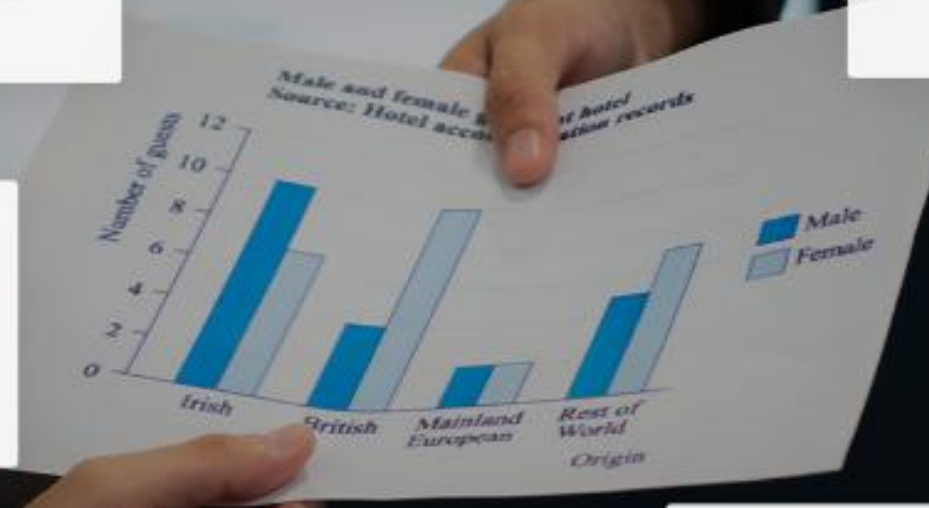
Data that does not fit into a strict structure but contains tags or other markers to separate data elements.

Unstructured Data

Data without a predefined format, making it challenging to organize and analyze efficiently.

Integration Challenges

Difficulties encountered when combining different types of data into a unified format for analysis and decision-making.



Analysis Complexity

The complexity in processing and extracting insights from diverse data forms due to their varied nature and formats.

Data Examples

Illustrative instances including text, images, video files, and sensor data to showcase the breadth of data variability.

Veracity

Data Quality

Veracity focuses on the accuracy, quality, and trustworthiness of the data.

Data Cleaning

Addressing data inconsistencies and inaccuracies is crucial for ensuring veracity in big data analytics.

Reliability

Veracity emphasizes the reliability of data for making informed decisions.

Value: Transforming Data into Insights

Unlocking the Power of Data

1

Turning Raw Data into Insights

The process of converting raw data into meaningful information that drives informed decision-making and strategic actions.

2

Data-Driven Decision-Making

Leveraging data analytics to make decisions based on evidence, trends, and statistical insights rather than intuition or guesswork.

3

Enhancing Business Intelligence

Utilizing data analytics tools and techniques to gain a competitive edge, improve operational efficiency, and enhance overall performance.

4

Personalized Marketing Strategies

Crafting targeted marketing campaigns by analyzing customer data to deliver personalized experiences and drive customer engagement and loyalty.

Variability: The Changing Nature of Data

Understanding the Evolution of Data Patterns

Changes in Data Patterns

Data patterns constantly evolve, reflecting shifts in trends, behaviors, and preferences.

Adapting to Data Landscape

Organizations must adjust strategies to align with the ever-changing data environment for sustainable growth.

Examples of Change

Illustrating how customer preferences transform over time showcases the dynamic nature of data.

The Interconnection of the 6 V's

Exploring the Complex Relationships Among Data V's

Volume vs. Velocity

1

The increase in Volume often influences the Velocity of data processing, requiring faster analytics to manage large datasets efficiently.

Variety vs. Veracity

3

Diverse data sources increase the challenge of ensuring data Veracity, emphasizing the importance of data quality and accuracy.

Value vs. Variability

5

The Value extracted from data analysis is subject to Variability based on the quality and consistency of data sources, affecting decision-making outcomes.

Velocity vs. Variety

2

Higher Velocity of data flow usually correlates with a greater Variety of data sources, leading to the need for versatile data processing methods.

Veracity vs. Value

4

Data Veracity directly impacts the Value derived from data analysis, emphasizing the need for reliable, trustworthy data.

Scalability and Parallel Processing in Big Data Analytics

Big Data needs

- Processing of large data volume
- Intensive computations
- Scalability enables increase or decrease in the capacity of data storage, processing and analytics, as per the complexity of computations and volume of data

How to Tackle the Big Data Performance Challenge

Three approaches to improving performance by orders of magnitude are:

- *Scale up* the computing resources on a node, via parallel processing & faster memory/storage
- *Scale out* the computing to distributed nodes in a cluster/cloud or at the edge
- *Scale down* the amount of data processed or the resources needed to perform the processing

Types of Scalability

Vertical Scalability (Scaling Up): Adding more power (CPU, RAM) to an existing machine.

Horizontal Scalability (Scaling Out): Adding more machines to handle the workload.

Importance: Essential for managing increasing data volumes efficiently.

Vertical Scalability (Scaling up)

- Means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities
- Solve problems of greater complexities by scaling up
- Architecture aware algorithm design

Vertical Scalability (Scaling up)

- Means designing the algorithm according to the architecture that uses resources efficiently

- For example, **x TB** of data take **time t** for processing, code size with increasing complexity increase by **factor n** , then scaling up means that processing takes equal, less or much less than **(n × t) for x TB**.

Horizontal Scalability (Scaling Out)

- Horizontal scalability means increasing the number of systems working in coherence and scaling out the workload

- Processing different datasets of a large dataset by increasing number of systems running in parallel.

Horizontal Scalability (Scaling Out)

- Scaling out means using more resources and distributing the processing and storage tasks in parallel

- If **r resources** in a system process **X TB** of data in **time t** , then (p × X) TB on **p parallel** distributed nodes such that the time taken up remains t or is slightly more than t

Scale Down Now

- **Several algorithms Exists**

  Reduce data size, data dimensionality, memory needed, etc

  E.g., Twitter's (X) open source Summingbird toolkit

  Hyperlog– number of unique users who perform a certain action; followers-of-followers

  CountMin Sketch – number of times each query issued to Twitter (X) search in a span of time; building histograms

  Bloom Filters – keep track of users who have been exposed to an event to avoid duplicate impressions (10^8 events/day for 10^8 users)

High Performance Capabilities

- Simple execution model – scalable, distributed, and parallel computing

- Deploy ‘Massively Parallel Processing’ (MPP) Platforms (MPPs), cloud, grid, clusters, and distributed computing software

Parallel Processing Overview

- The simultaneous processing of multiple tasks to increase computational speed.
- ***Comparison with Serial Processing:*** Serial processing handles one task at a time, whereas parallel processing handles multiple tasks concurrently.
- **Benefits:**

➢ Faster data processing

➢ Efficient resource utilization

➢ Improved performance for complex computations

Scalability Techniques : Components

Sharding: Dividing a database into smaller, more manageable pieces.

Partitioning: Splitting data into distinct, independent parts.

Load Balancing: Distributing workloads evenly across multiple servers.

Caching : storing frequently accessed data in a cache to reduce the need to access the original source of the data.

Microservices Architecture : Microservices architecture involves breaking down a monolithic application into smaller, more independent services.

Content Delivery Networks (CDNs): caching and delivering content from servers that are geographically closer to users, reducing latency and improving performance

Parallel Processing Frameworks

Apache Hadoop:

- Features: Distributed storage (HDFS), MapReduce processing model.
- Use Cases: Batch processing, large-scale data storage.

Apache Spark:

- Features: In-memory processing, real-time data processing.
- Use Cases: Real-time analytics, machine learning.

Apache Flink:

- Features: Stream processing, fault tolerance.
- Use Cases: Event-driven applications, real-time data processing.

Comparison: Performance, ease of use, community support.

Distributed File Systems

- Enable storage and access to large datasets across multiple machines.

Hadoop Distributed File System (HDFS):

- Architecture: Master/slave, data blocks distributed across nodes.
- Benefits: Fault tolerance, high throughput.

- Challenges: Data replication overhead, latency in accessing distributed data.

Data Partitioning and Distribution

Techniques:

- **Range Partitioning:** Dividing data into ranges based on key values.
- **Hash Partitioning:** Distributing data based on hash values of keys.
- **Round-Robin Partitioning:** Evenly distributing data across partitions.

Strategies:

- Data Locality: Moving computation to where the data resides.
- Replication: Storing multiple copies of data across nodes.

Data Storage and Analysis

2. Data Storage and Analysis

- **Data storage and Management : Traditional Systems**
  - Structured or Semi Structured Data
  - SQL
  - Large Data Storage using RDBMS
  - Distributed Database Management Systems
  - In-Memory Column Format Databases
  - In-Memory Row Format Databases
  - Enterprise Data Store Server and Data Warehouse

- **Big Data Storage**
  - No SQL
  - Coexistence with Data sources

- **Big Data Platform**

Data storage and Management : Traditional Systems

Structured Data

- RDBMS data, such as MySQL DB2, enterprise server and data warehouse
- SQL– a language for managing the RDBMS Relational database examples are MySQL PostGreSQL Oracle database, Informix, IBM DB2 and Microsoft SQL server

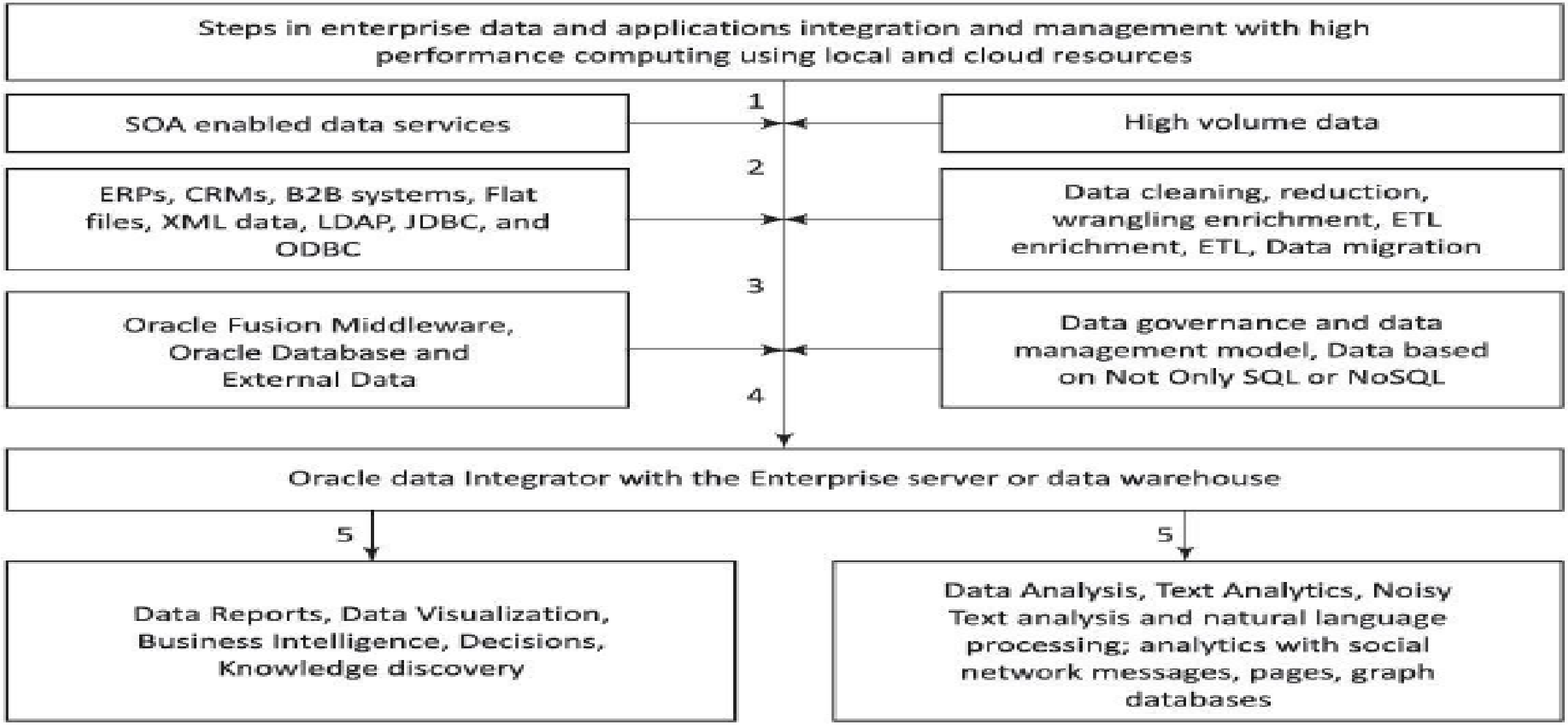
Semi-Structured Data

- XML and JSON
- A comma-separated values (CSV) file

Data storage and Management : Traditional Systems

- **Enterprise Data**
  - Enterprise Data-Store Server
  - Data Warehouse
  - Enterprise data warehouse store the databases, and data stores after integration, using tools from number of sources

Enterprise Data Integration and Management with Big Data for HPC



Source : Big Data Analytics, Rajkamal

Big Data Storage

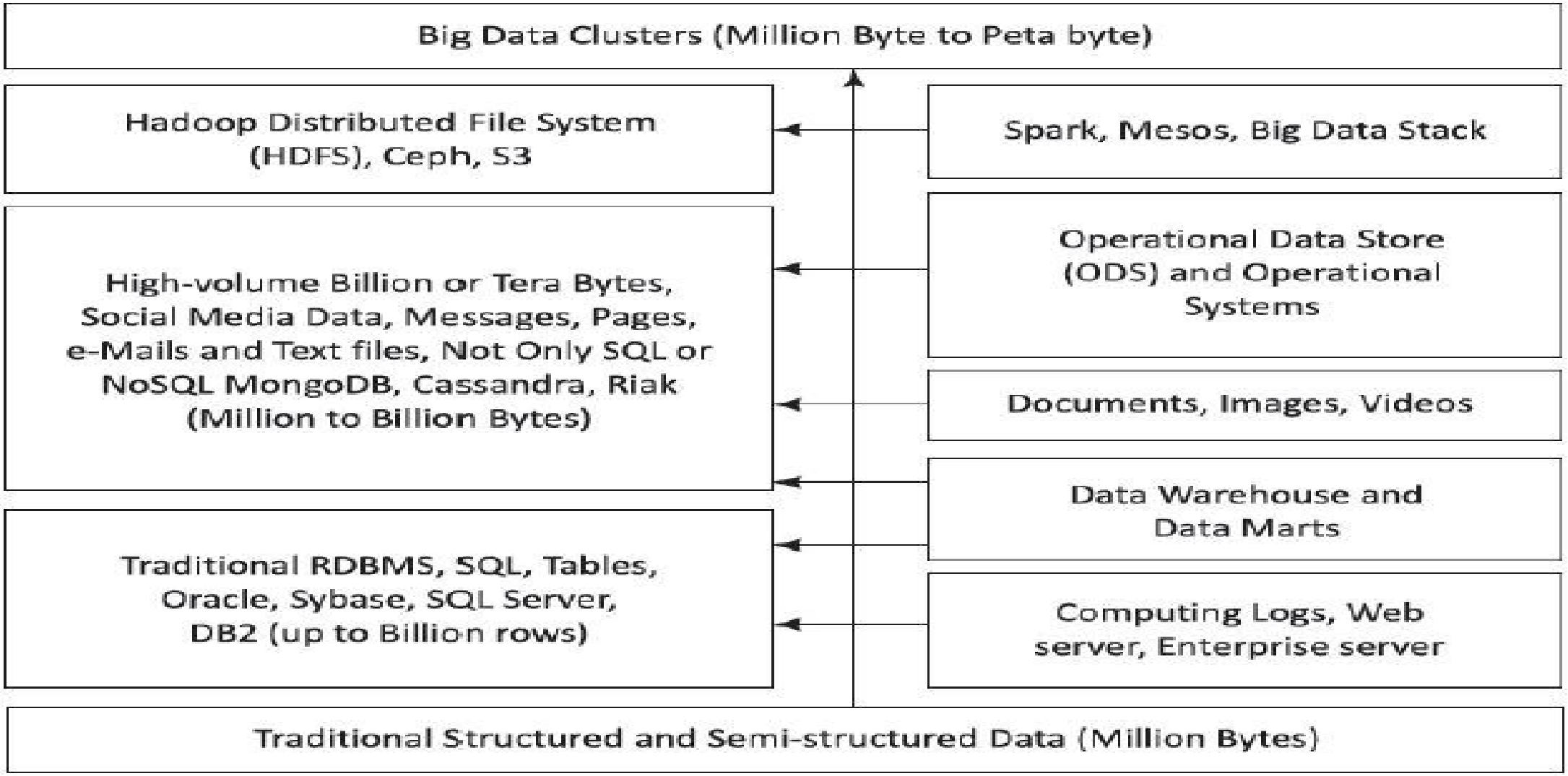
- Big Data Store uses NoSQL. NOSQL stands for No SQL or Not Only SQL
- NoSQL databases considered as semi structured data
- The stores do not integrate with applications using SQL
- Features of NoSQL – Many , CAP theorem
- Refer the Table for ***Various Data Sources and examples of usages and tools in the next slides***

Various Data Sources and examples of usages and tools

Data Source	Examples of Usages	Example of Tools
Relational databases	Managing business applications involving structured data	Microsoft Access, Oracle, IBM DB2, SQL Server, MySQL, PostgreSQL Composite, SQL on Hadoop [HPE (Hewlett Packard Enterprise) Vertica, IBM BigSQL, Microsoft Polybase, Oracle Big Data SQL]
Analysis databases (MPP, columnar, In-memory)	High performance queries and analytics	Sybase IQ, Kognitio, Terradata, Netezza, Vertica, ParAccel, ParStream, Infobright, Vectorwise,
NoSQL databases (Key-value pairs, Columnar format, documents, Objects, graph)	Key-value pairs, fast read/write using collections of name-value pairs for storing any type of data; Columnar format, documents, objects, graph DBs and DSs	Key-value pair databases: Riak DS (Data Store), OrientDB, Column format databases (HBase, Cassandra), Document oriented databases: CouchDB, MongoDB; Graph databases (Neo4j, Tetan)

Hadoop clusters	Ability to process large data sets across a distributed computing environment	Cloudera, Apache HDFS
Web applications	Access to data generated from web applications	Google Analytics, Twitter
Cloud data	Elastic scalable outsourced databases, and data administration services	Amazon Web Services, Rackspace, GoogleSQL
Individual data	Individual productivity	MS Excel, CSV, TLV, JSON, MIME type
Multidimensional	Well-defined bounded exploration especially popular for financial applications	Microsoft SQL Server Analysis Services
Social media data	Text data, images, videos	Twitter, LinkedIn

Coexistence of various Data sources



Source : Big Data Analytics, Rajkamal

Big Data Platform

Big Data Platform

- Supports Large datasets and volume of Data
- Data Generated at **Higher Velocity, more varieties, Higher Veracity**
- **Provisioning needed for**
  - *Storage, Processing and Analytics*
  - *Developing, Deploying, Operating and Managing Big Data Environment*
  - *Reducing the Complexity and integration of Applications*
  - *Custom Development, Querying*
  - *Traditional as well as Big Data Techniques*

Big Data Platform Data management, Storage and Analytics

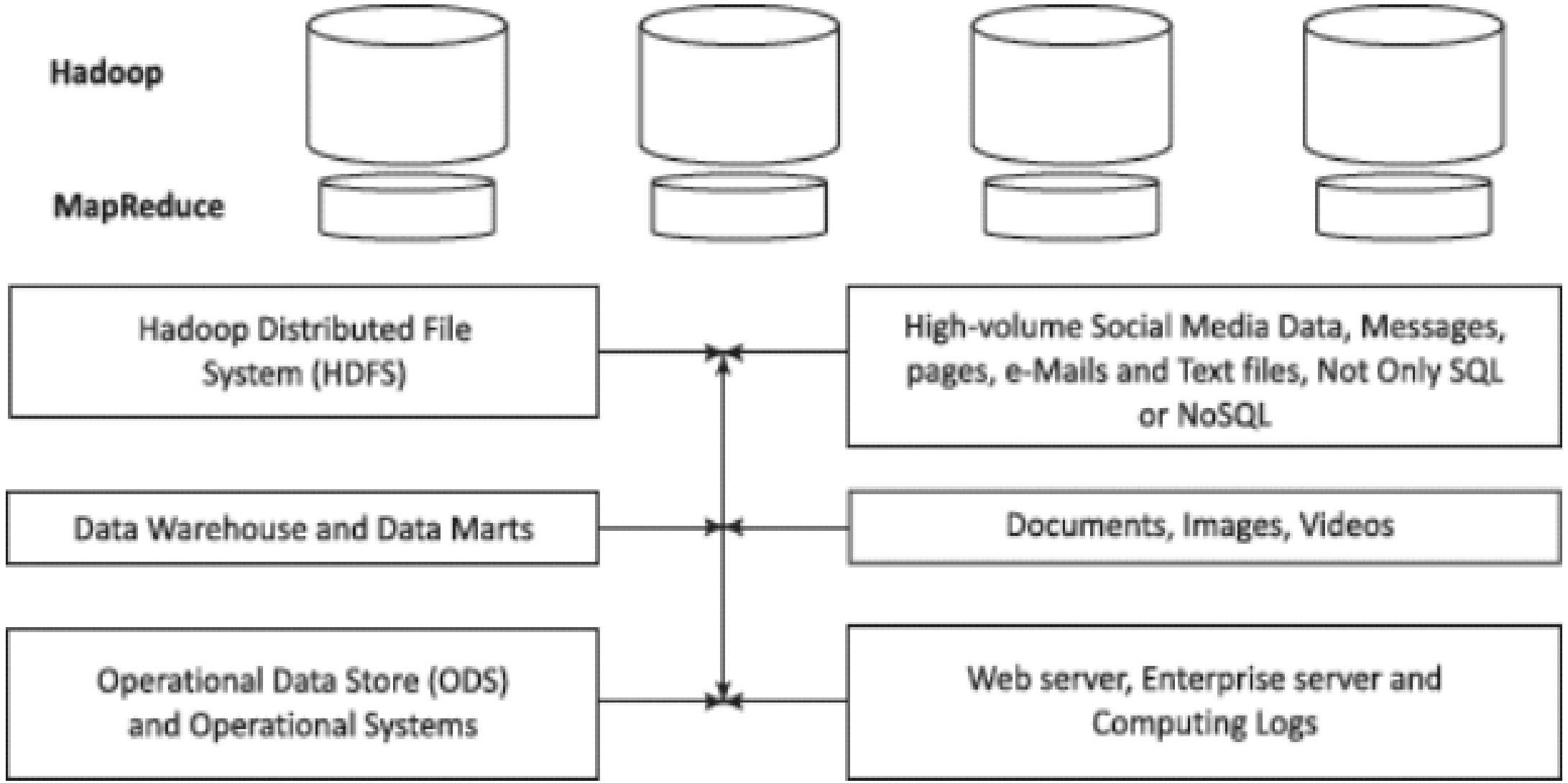
Following are requirements:-

- New innovative non-traditional methods of storage, processing and analytics
- Distributed Data Stores
- Creating scalable as well as elastic virtualized platform (cloud computing)
- Huge volume of Data Stores
- Massive parallelism
- High speed networks
- High performance processing, optimization and tuning

Big Data Platform Data management, Storage and Analytics (con..)

- Data management model based on Not Only SQL or NoSQL
- In-memory data column-formats transactions processing or dual inmemory data columns as well as row formats for OLAP and OLTP
- Data retrieval, mining, reporting, visualization and analytics
- Graph databases to enable analytics with social network messages, pages and data analytics
- Machine learning or other approaches
- Big data sources: Data storages, data warehouse, Oracle Big Data, MongoDB NoSQL, Cassandra NoSQL
- Data sources: Sensors, Audit trail of Financial transactions data, external data such as Web, Social Media, weather data, health records data.

Hadoop based Big Data environment



Source : Big Data Analytics, Rajkamal

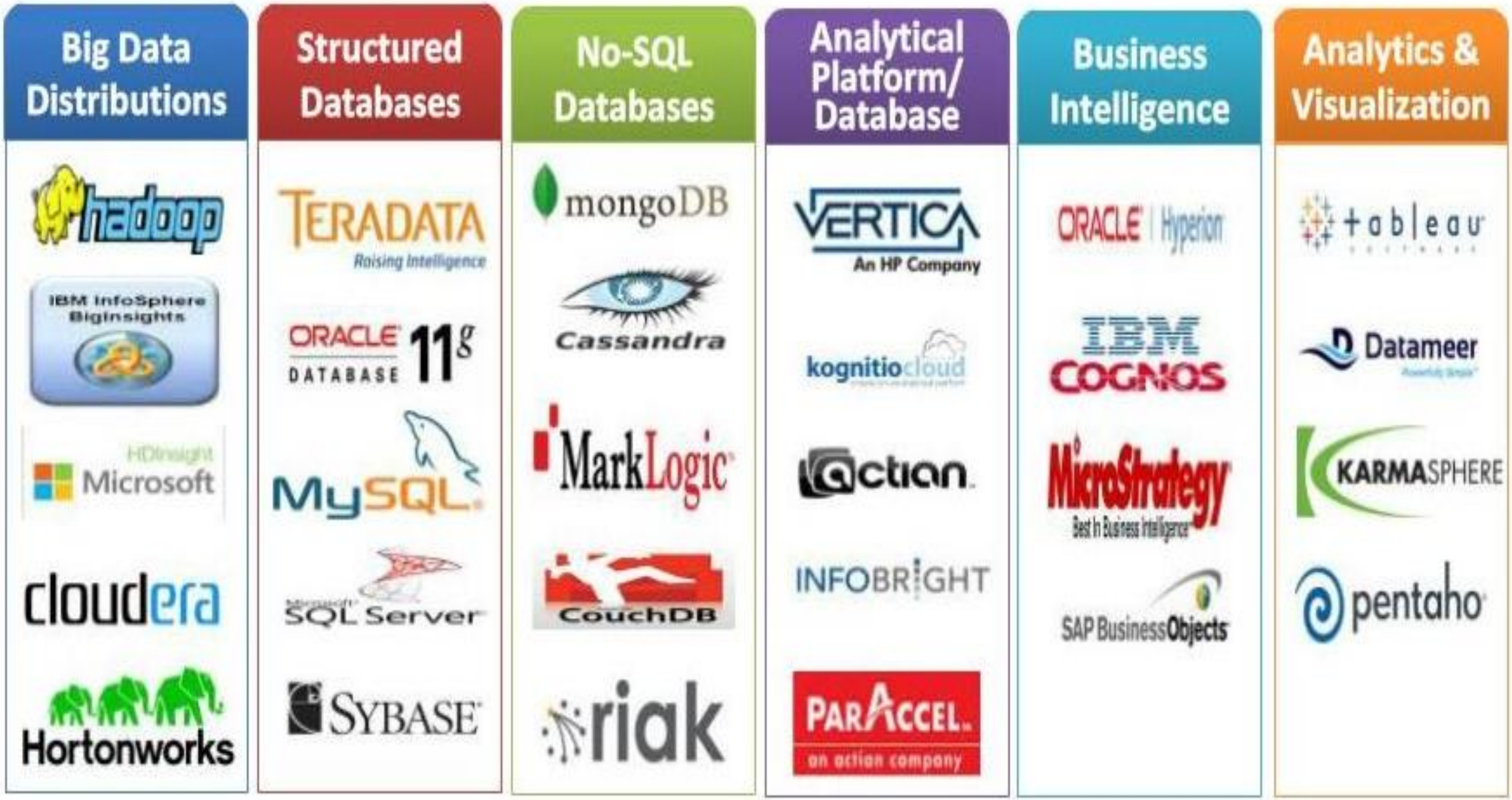
Big data Stack

- A stack contains set of Software Components and Data store units

- **How it is used?**
  - For Applications, Machine Learning Algorithms, Analytics and Visualization Tools.
  - Use Big Data Stack (BDS) at a cloud service, such as Amazon EC2, Azure or private cloud. The stack uses cluster of high performance machines

Tools for Big Data Environment

TYPES	EXAMPLES
MapReduce	Hadoop, Apache Hive, Pig, Cascading, Cascalog, mrjob (Python Mapreduce Library),Apache S4, MapR, Apple Acunu, Apache Flume Apache Kafka
NoSQL Databases	MongoDB, Apache CouchDB, Apache Cassandra, Aerospile, Apache Hbase, Hypertable
Processing	Spark,IBM Big Sheets,PySpark, R, Yahoo Pipes,Amazon Mechanical Turk, Datameter, Apache Solr/Lucene, Elastic Search
Servers	Amazon EC2, S3, GoogleQuery, GoogleAppEngine, AWS Elastic Beanstalk, Salesforce, Heroku
Storage	HDFS, Amazon S3, Mesos

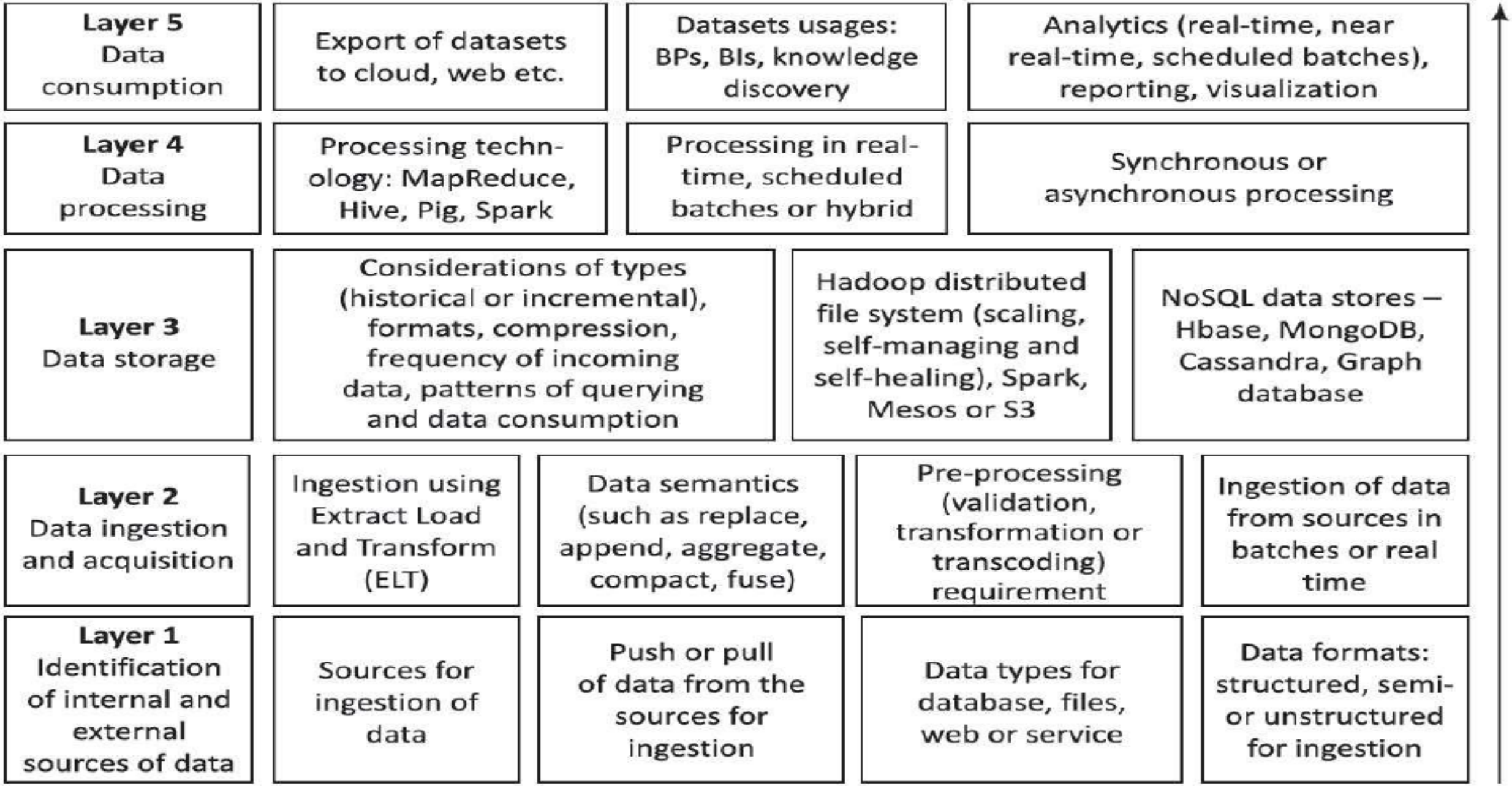

Big Data Distributions
hadoop
IBM InfoSphere BigInsights
HDInsight
Microsoft
cloudera
Hortonworks
Structured Databases
TERADATA
Raising Intelligence
ORACLE 11g DATABASE
MySQL
SQL Server
SYBASE
No-SQL Databases
mongoDB
Cassandra
MarkLogic
CouchDB
riak
Analytical Platform/ Database
VERTICA
An HP Company
kognitiocloud
action
INFOBRIGHT
PARACCEL
an action company
Business Intelligence
ORACLE Hyperion
IBM COGNOS
MicroStrategy
Best In Business Intelligence
SAP BusinessObjects
Analytics & Visualization
tableau
Datameer
KARMASPHERE
pentaho

Design Layers in Data Processing Architecture

Big Data Architecture

- Big Data architecture is the logical and/or physical layout/structure of how Big Data will be stored, accessed and managed within a Big Data or IT environment

- Logically defines how Big Data solution will work, the core components (hardware, database, software, storage) used, flow of information, security and more



Source : Big Data Analytics, Rajkamal

Types of Data Analytics

Types of Data Analytics

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

Descriptive Analytics

- Definition: Summarizes past data to understand what happened
- Key techniques: data aggregation, data mining
- **Examples:**

- Sales Reports: Analyzing historical sales data to identify trends

- Healthcare: Summarizing patient data to understand disease prevalence

Descriptive Analytics

Example: Retail Sector

- Use of real-time sales dashboards
- Analysis of customer purchase history for personalized marketing

Example: Hospital Resource Optimization during Flu Season

-Tech Stack & Tools: Real-time dashboards (Power BI), SQL + Python for historical data aggregation, and Apache Kafka for live updates on bed usage.

- Outcomes: 17% faster bed turnaround, 22% reduction in patient wait times, and 90% fewer medicine stockouts—supporting data-driven operational decisions.

Diagnostic Analytics

Definition: Investigates why something happened

Key techniques: drill-down, data discovery

Examples:

- Sales Decline: Identifying reasons for a drop in sales
- Finance : Investigating causes of Financial Services dropdown

Diagnostic Analytics

Example: Financial Services

- Using diagnostic analytics to understand reasons behind market fluctuations

- Use Case: A fintech company detects a sudden drop in credit card transaction volume for a specific customer segment in Q2 2025.

Analysis Performed:

- Drill-down by region, age group, and income level
- Correlation analysis with interest rate hikes and seasonal purchasing trends
- Segmentation to identify affected cohorts (e.g., Gen Z in Tier 2 cities)
- Insight Gained: The decline was strongly linked to recent RBI policy changes and reduced digital spending in select regions—enabling targeted promotions and revised credit risk models.

Predictive Analytics

Definition: Forecasts future trends based on historical data

Key techniques: statistical modeling, machine learning

Examples:

- Sales Forecasting: Predicting future sales based on past data
- Healthcare: Predicting disease outbreaks

Predictive Analytics

Use Case: Forecast optimal fishing zones and catch volumes using satellite data (SST, chlorophyll), vessel GPS, and historical catch records.

Models Used: MLs to identify high-probability catch areas based on oceanographic and environmental factors.

Impact: Achieved up to 88% prediction accuracy, 30% fuel savings, and 35% increase in catch efficiency, enabling sustainable and data-driven fishing operations.

Prescriptive Analytics

Definition: Recommends actions based on data

Key techniques: optimization, simulation algorithms

Examples:

- Supply Chain Optimization: Recommending optimal inventory levels
- Healthcare: Personalized treatment plans based on patient data

Prescriptive Analytics

Example: Logistics

- Using AI to optimize delivery routes in real-time
- Recommending inventory restocking based on predictive models

Comparison of Data Analytics Types

Metric	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Purpose	Summarizes past data to understand what happened	Investigates why something happened	Forecasts future trends based on historical data	Recommends actions based on data
Techniques	Data aggregation, data mining	Drill-down, data discovery	Statistical modeling, machine learning	Optimization, simulation algorithms
Example1	Sales reports, healthcare data summaries	Identifying reasons for sales decline, investigating disease outbreak	Sales forecasting, predicting disease outbreaks	Supply chain optimization, personalized treatment plans
Example2	Real-time sales dashboards, customer purchase history analysis	Financial services market fluctuations analysis, customer churn analysis	Machine learning for patient readmissions, forecasting infectious diseases	AI for optimizing delivery routes, recommending inventory restocking based on models