BITS Pilani

Module 9 Part 3
Big Data technologies

Harvinder S Jabbal
SSZG653 Software Architectures

Contents

1. Big data
2. Hadoop, HDFS, Map-Reduce
3. Analytics & Real time analytics
4. In-Memory database
5. NoSQL databases

Big data & Analytics

Wikipedia defines ***"Big Data"*** as a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

History of Big Data

Lots of data got created due to

- Proliferation of Internet
- Social media
- eCommerce

Before we begin, let us understand the scale of data we deal with

Unit	Power of 10
Mega byte	6
Giga	9
Tera	12
Peta	15

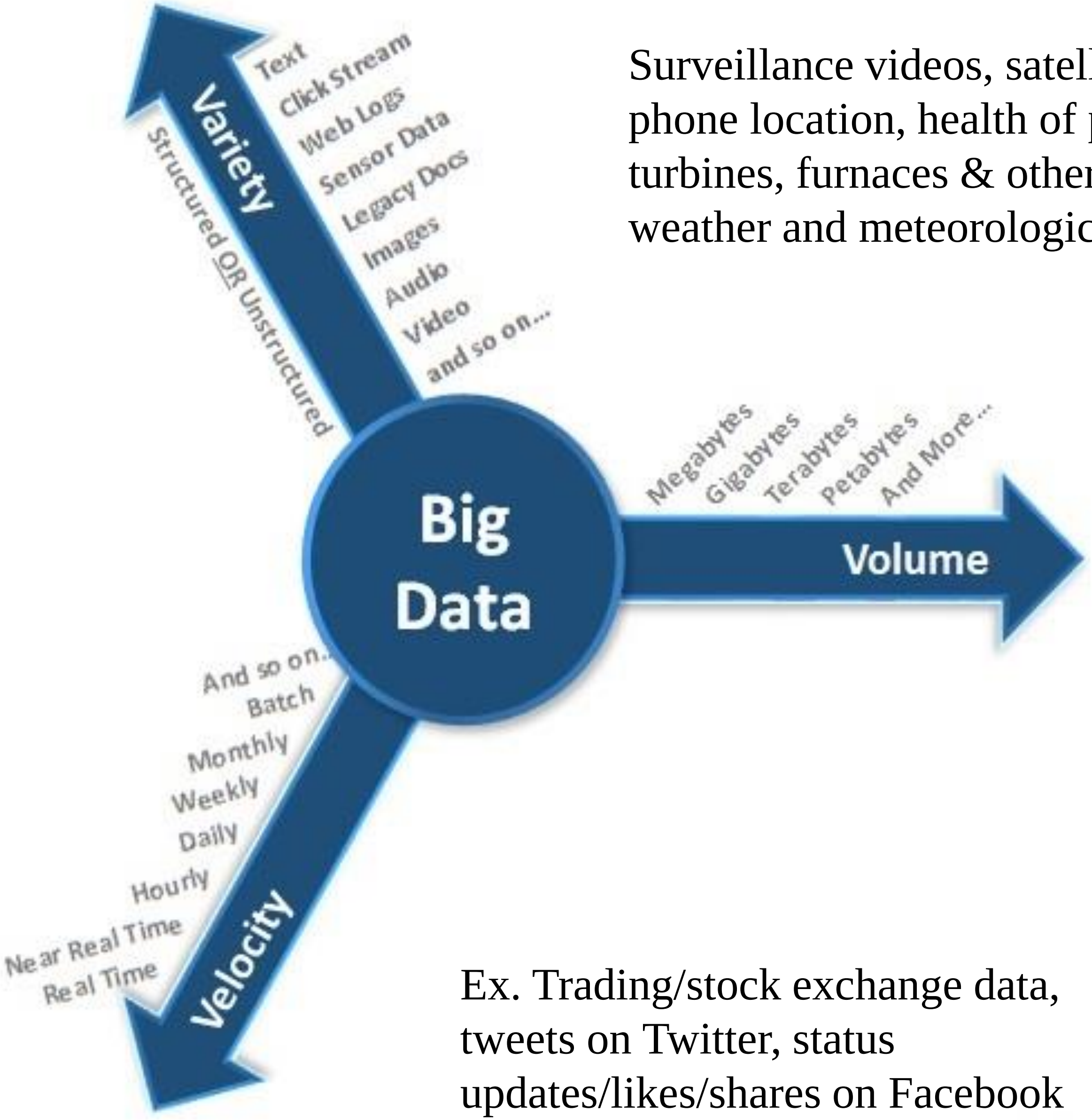
Big Data Statistics

- Volume of data
  - 500+ new websites are created every minute of the day
  - 100 Terabytes of data is uploaded to Facebook every day
  - Twitter generates 12 Terabytes of data every day
  - YouTube users upload 48 hours of new video content **every minute** of the day

- Processing
  - Decoding of the human genome used to take 10 years. Now it can be done in 7 days
  - Facebook Stores, Processes, and Analyzes more than 30 Petabytes of user generated data
  - LinkedIn processes and mines Petabytes of user data to power the "People You May Know" feature

Source: Wikibon - A Comprehensive List of Big Data Statistics

Characteristics of Big data



Surveillance videos, satellites images, cell phone location, health of power station turbines, furnaces & other industrial machinery, weather and meteorological data, click stream

Archived data:
Patient records, Scanned copies of agreements, records of ex-employees/completed projects, banking transactions older than the compliance regulations

Ex. Trading/stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook

Use of big data

- Recommend cancer medication based on what worked well in similar situations for other patients
- Weather prediction for fishermen, farmers
- Predict equipment malfunctioning in large nuclear power plant, chemical plants, etc.
- Credit card fraud detection

Example of handling big data

Google search

Do you know how Google Search works?

3 steps:

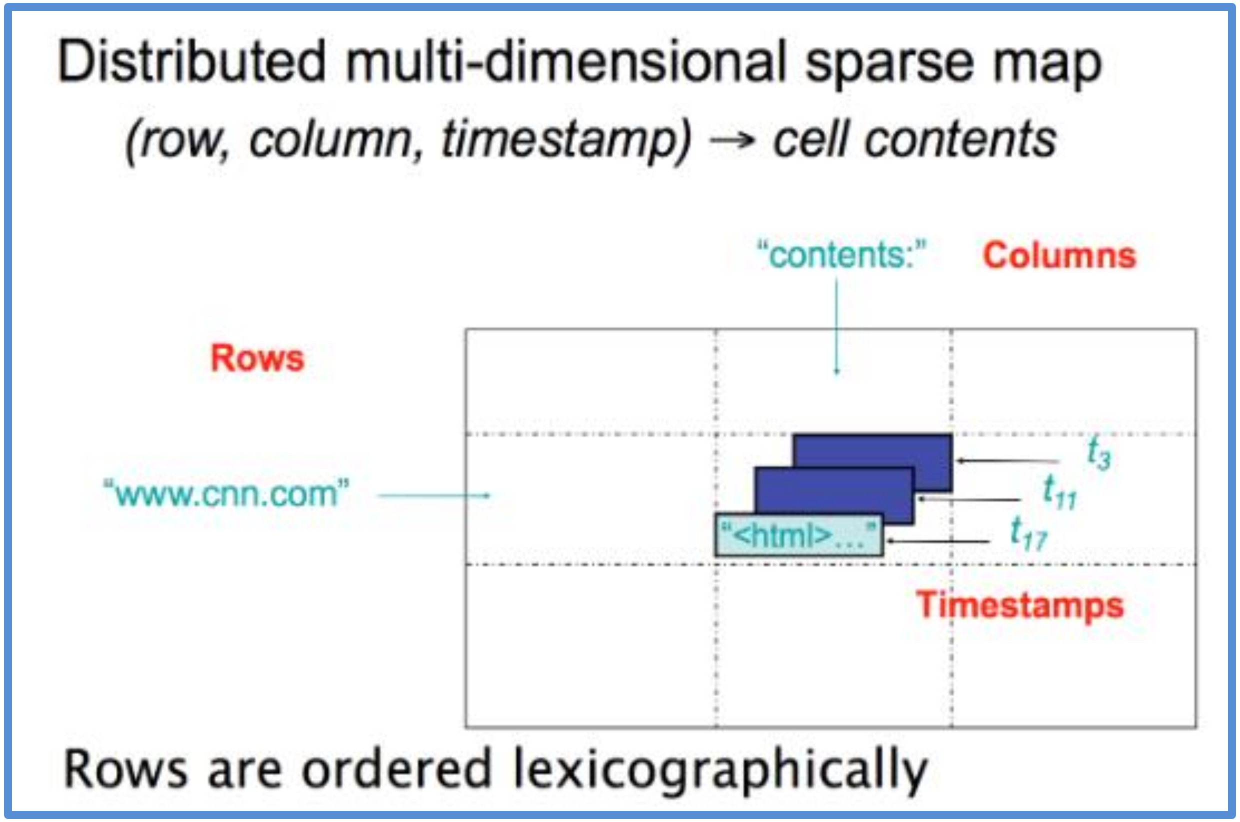
- **Catalog & index**: Even before you search, Google goes through various web sites and linked web sites (crawling) and catalogs all the web sites. This runs into several Tera bytes.

- **Understand & enrich your query**: Correct spelling mistakes, consider synonyms, etc.

- **Search:** When the search is requested, it uses this catalog to determine which web sites closely match the requirement. For this it uses a ‘Page Rank’ (named after Larry Page) algorithm which considers factors such as which page has max number of occurrences of the search string, how many other web sites refer to this web site, what is the reputation of the websites that refer to this website, etc.

Google – Big table

So Google invented a data storage structure called Big Table



Distributed multi-dimensional sparse map

(row, column, timestamp) → cell contents

Rows are ordered lexicographically

Google BigTable

Google BigTable is a wide table containing several attributes of a website;

Here are some attributes:

- The contents of Web page
- Anchor text*
- Websites referencing the page.
- Time stamp when the data was stored

Google BigTable is built on technologies like Google File System (GFS)

Google BigTable is used by applications such as Google Maps, Google Analytics, etc.

*Anchor text is the text that appears in Blue colour in search result. It contains a link to the website

Google Big Table

Features:

- Versioning of data,
- Compression,
- Distribution across servers,
- Fault tolerant,
- Fast access,
- Dynamic addition of servers,
- Load balancing

Google disclosed the design of Big table. Then came Hadoop Distributed File System (Yahoo) and serveral NoSQL databases

Use of Big Data

Banking and Financial Services

- Fraud Detection to detect the possible fraud or suspicious transactions in Accounts, Credit Cards, Debit Cards, and Insurance etc.

Retail

- Targeting customers with different discounts, coupons, and promotions etc. based on demographic data like gender, age group, location, occupation, dietary habits, buying patterns, and other information which can be useful to differentiate/categorize the customers.

Sentiment Analysis

- Organizations use the data from social media sites like Facebook, Twitter etc. to understand what customers are saying about the company, its products, and services.
- Words like “I like this phone”, “This food is too salty”, etc.. indicate sentiments
- This type of analysis is also performed to understand which companies, brands, services, or technologies people are talking about.

Use of Big Data …

Customer Service

- IT Services and BPO companies analyze the call records/logs to gain insights into customer complaints and feedback, call center executive response/ability to resolve the ticket, and to improve the overall quality of service.
- Call center data from telecommunications industries can be used to analyze the call records/logs and optimize the price, and calling plan, messaging plan, and data plans

Industrial equipment monitoring & alerting

- A large power plant or chemical factory has thousands of critical equipment that needs to be monitored
- The equipment data needs to be analysed to detect any malfunctioning or danger of accidents

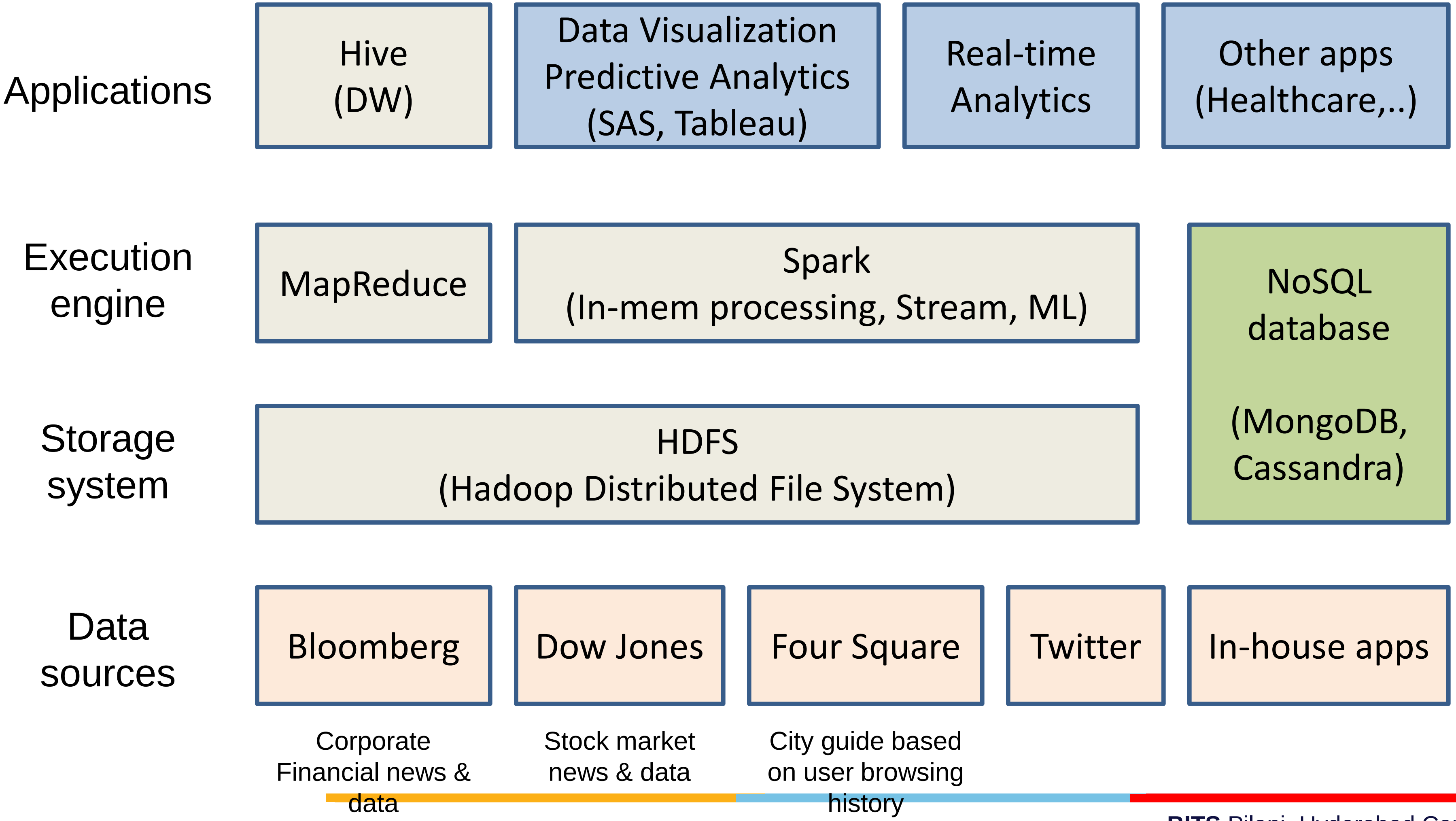
Weather forecasting

- Satellite data from remote sensing satellites need to be analysed at high speed to warn fishermen, farmers and public about potential cyclones, delayed monsoon, etc.

BigTable can be used to store…

- **Time-series data,** such as CPU and memory usage over time for multiple servers.
- **Marketing data,** such as purchase histories and customer preferences.
- **Financial data,** such as transaction histories, stock prices, and currency exchange rates.
- **Internet of Things data,** such as usage reports from energy meters and home appliances.
- **Graph data,** such as information about how users are connected to one another.

Big data architecture / Eco-system



Applications

- Hive (DW)
- Data Visualization Predictive Analytics (SAS, Tableau)
- Real-time Analytics
- Other apps (Healthcare,..)

Execution engine

- MapReduce
- Spark (In-mem processing, Stream, ML)

Storage system

- HDFS (Hadoop Distributed File System)

NoSQL database

(MongoDB, Cassandra)

Data sources

- Bloomberg – Corporate Financial news & data
- Dow Jones – Stock market news & data
- Four Square – City guide based on user browsing history
- Twitter
- In-house apps

Hadoop

Hadoop is an open source framework, from the Apache foundation, capable of processing large amounts of heterogeneous data sets in a distributed fashion across clusters of commodity computers and hardware using a simplified programming model.

Hadoop provides a reliable shared storage and analysis system.

Components of Hadoop

HDFS (Hadoop Distributed File System)

- HDFS offers a highly reliable and distributed storage, and ensures reliability, even on a commodity hardware, by replicating the data across multiple nodes.
- Unlike a regular file system, when data is pushed to HDFS, it will automatically split into multiple blocks (configurable parameter) and stores/replicates the data across various data nodes. This ensures high availability and fault tolerance.

MapReduce

- MapReduce offers an analysis system which can perform complex computations on large datasets.
- This component is responsible for performing all the computations and works by breaking down a large complex computation into multiple tasks and assigns those to individual worker/slave nodes and takes care of coordination and consolidation of results

Hadoop - HDFS

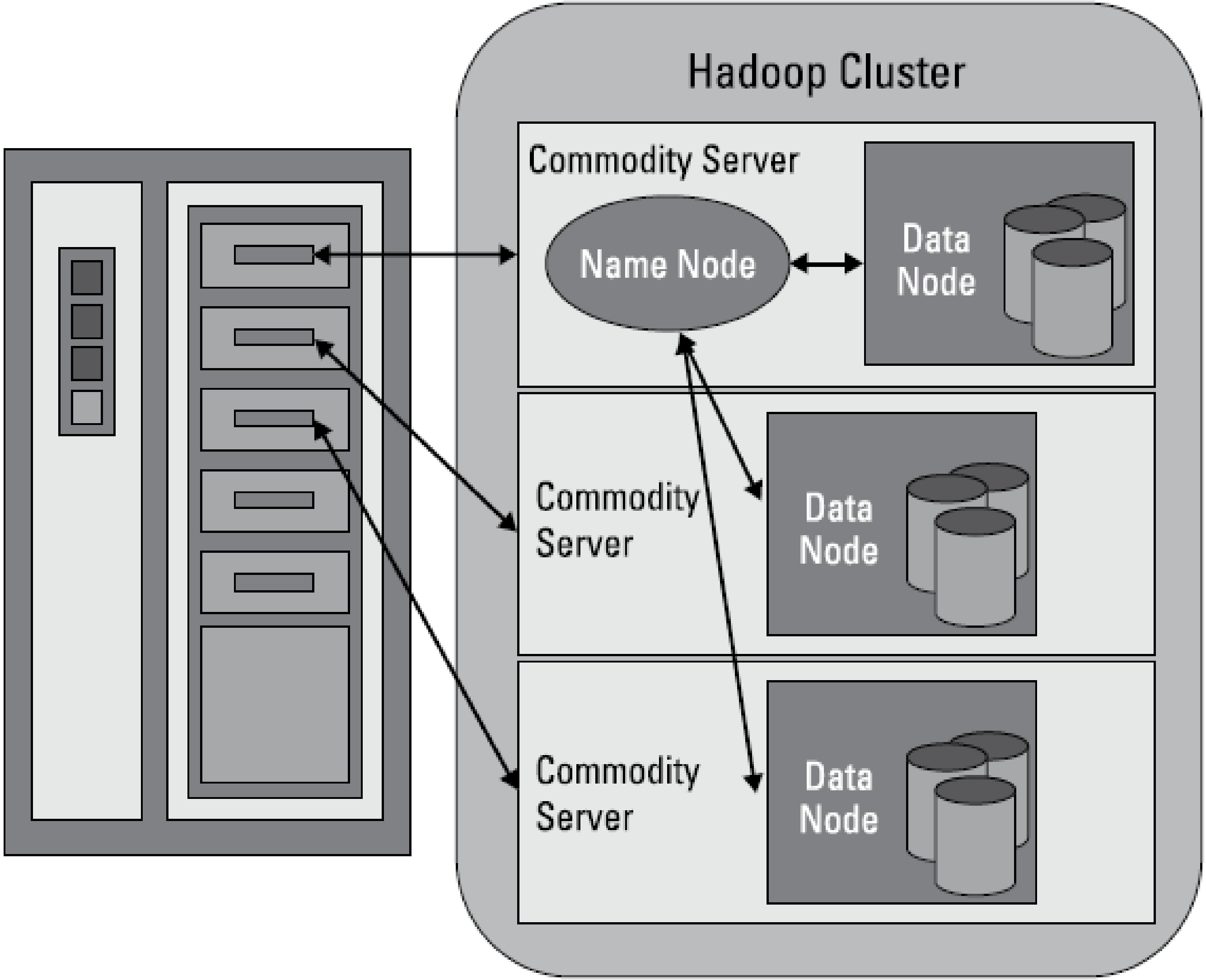


Figure 9-1: How a Hadoop cluster is mapped to hardware.

Ref: Big data for Dummies

Hadoop - HDFS

Data

- Large files are broken down into blocks (128 MB usually) and spread across Data nodes.
- Data blocks are replicated and Degree of replication can be adjusted

Meta data

- Name node stores meta data – data about files, distribution of data (which block is in which nodes), etc.
- For good performance, all the metadata is loaded into the physical memory of the NameNode server.

Features

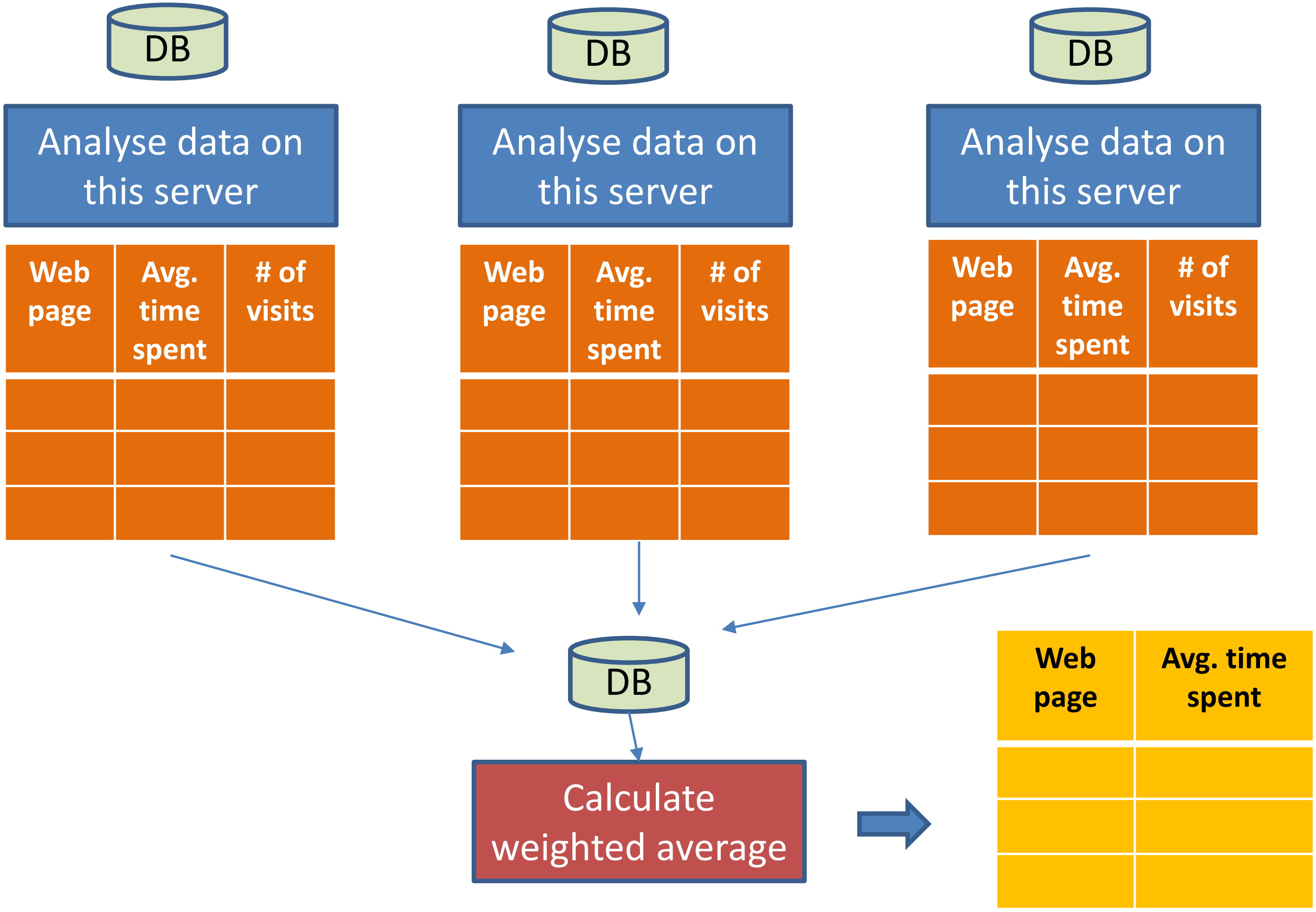
- Data nodes provide Heart beat messages to Name node
- Supports data pipelines. A connection between data nodes to move data from one node to another
- Rebalancer: Balances distribution of data

Map-Reduce pattern

- Used to analyse vast amount of data

- Suppose we keep track of every click of the user on a web site and store these details in a database
- Let us say we want to find out the average time spent by users on each web page of the web site, across thousands of users who visited the web site in the last 30 days
- How can we speed up the analysis?

Map-Reduce pattern: Example



DB

Analyse data on this server

Web page	Avg. time spent	# of visits

DB

Analyse data on this server

Web page	Avg. time spent	# of visits

DB

Analyse data on this server

Web page	Avg. time spent	# of visits

DB

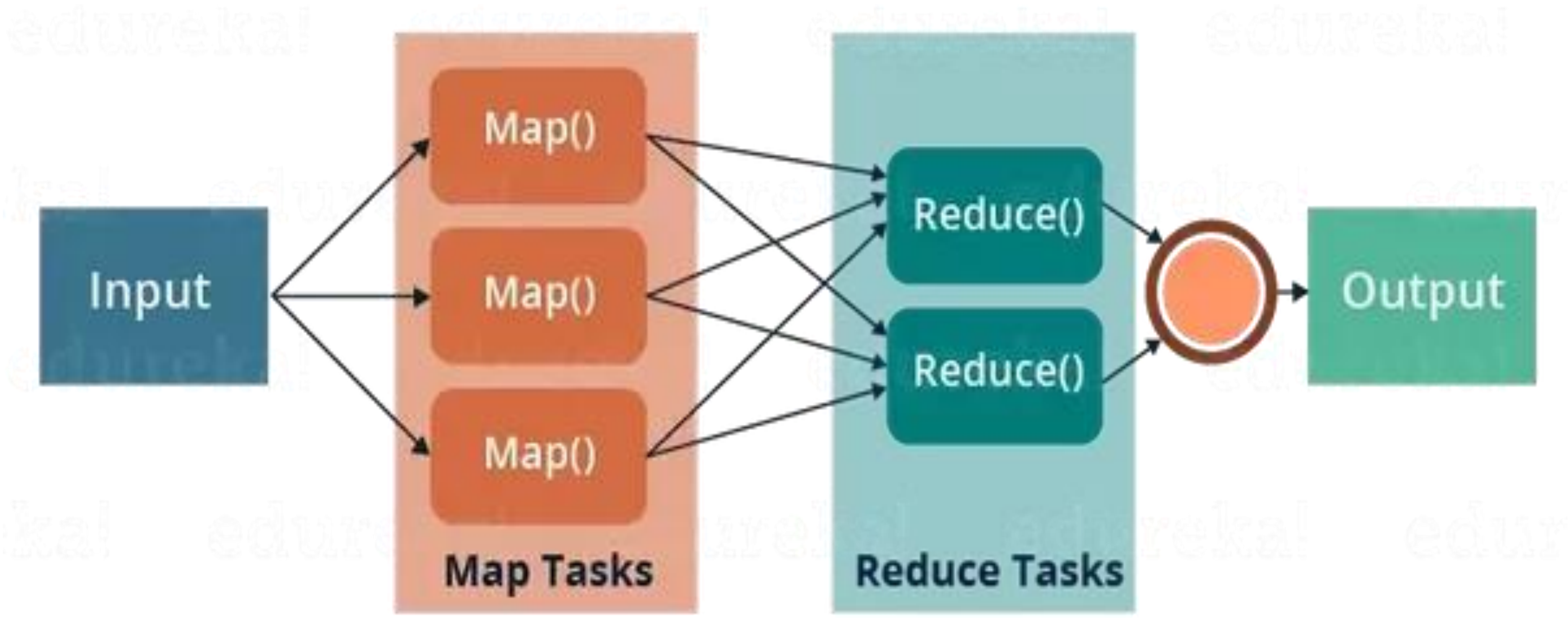
Calculate weighted average

Web page	Avg. time spent

Map-Reduce pattern

- Executes in parallel
- Leads to low latency & high availability
- Map performs extract & transform and produces <Key, Value> instances
- Reduce summarizes transformed data

Map Reduce pattern



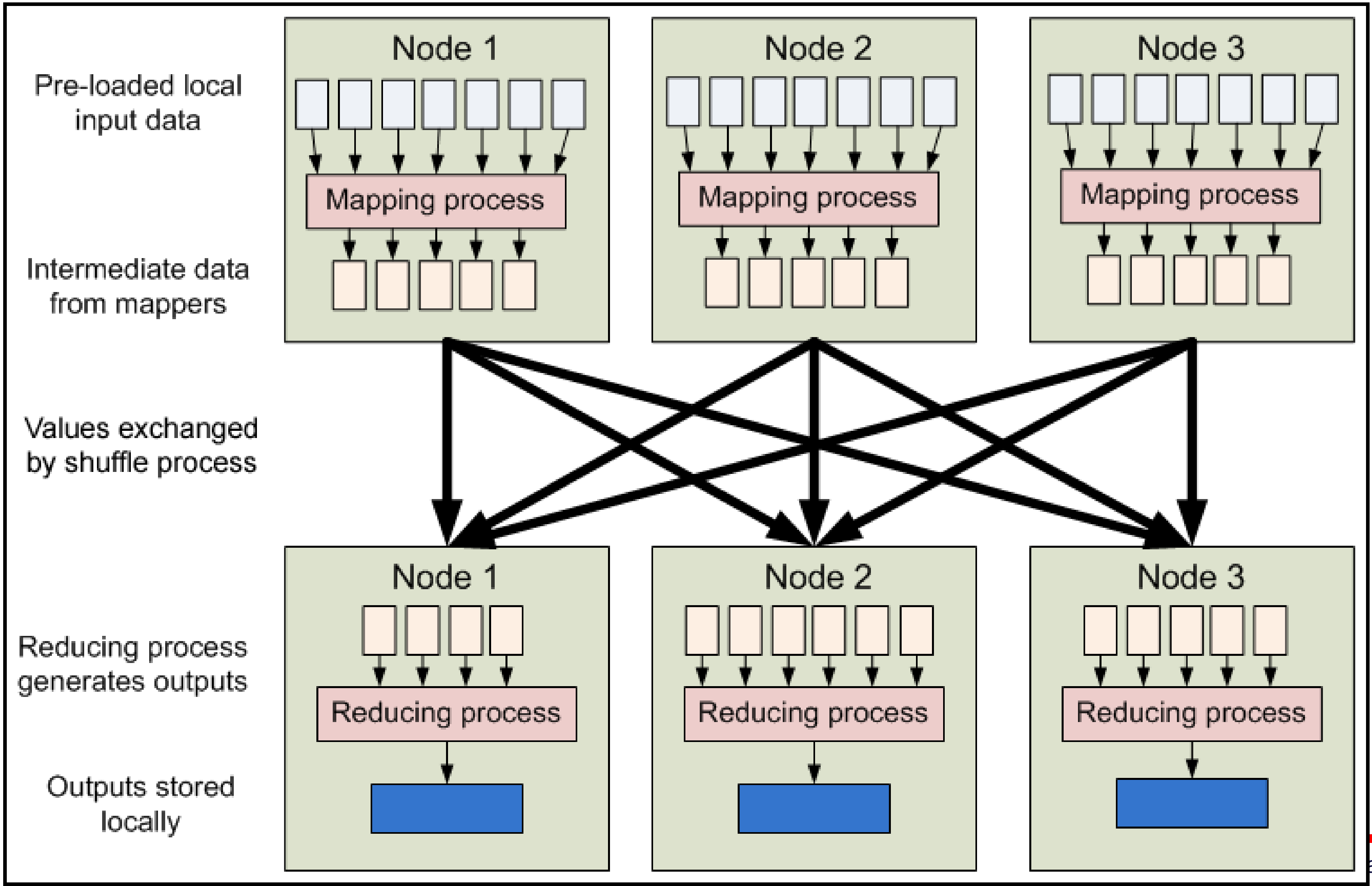
Source: https://www.quora.com/What-is-the-relationship-between-MapReduce-and-Hadoop

Map-Reduce pattern

- **Example**: Determine the average time (duration) spent by users on different web pages
- Step 1: **Map** processes data on each node are outputs <Web page, (Avg time, # of users)>
- Step 2: **Reduce** produces <Web page, weighted Avg time>

Map – Reduce pattern

Example: Determine the average duration spent by users on different web pages



Experience Sharing

Have you come across systems that use this pattern?

Hadoop - HDFS

Features

- Can store peta bytes of data
- Distributed
- Replicated
- Fault tolerant (self healing)

Using Hadoop

When to Use Hadoop (Hadoop Use Cases)

Hadoop can be used in various scenarios including some of the following:

- Analytics
- Search
- Data Retention
- Log file processing
- Analysis of Text, Image, Audio, & Video content
- Recommendation systems like in E-Commerce Websites

When Not to Use Hadoop

There are few scenarios in which Hadoop is not the right fit. Following are some of them:

- Low-latency or near real-time data access.
- If you have a large number of small files to be processed. This is due to the way Hadoop works. Namenode holds the file system metadata in memory and as the number of files increases, the amount of memory required to hold the metadata increases.
- Multiple writes scenario or scenarios requiring arbitrary writes or writes between the files.

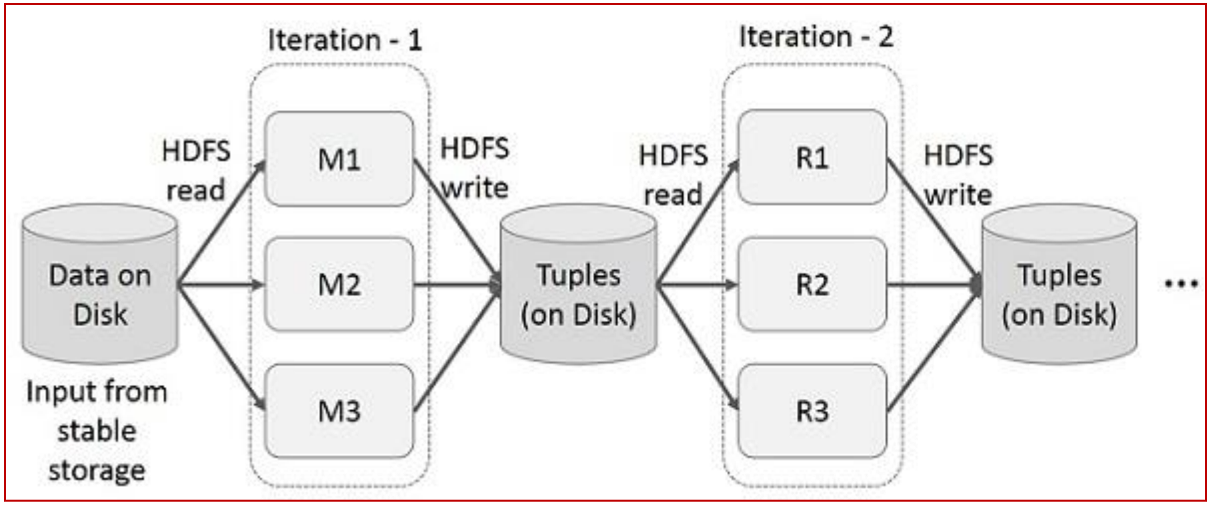
Hadoop may still be slower for some use cases

- Sometimes we require even faster processing than Map-Reduce, for example in real time fraud detection

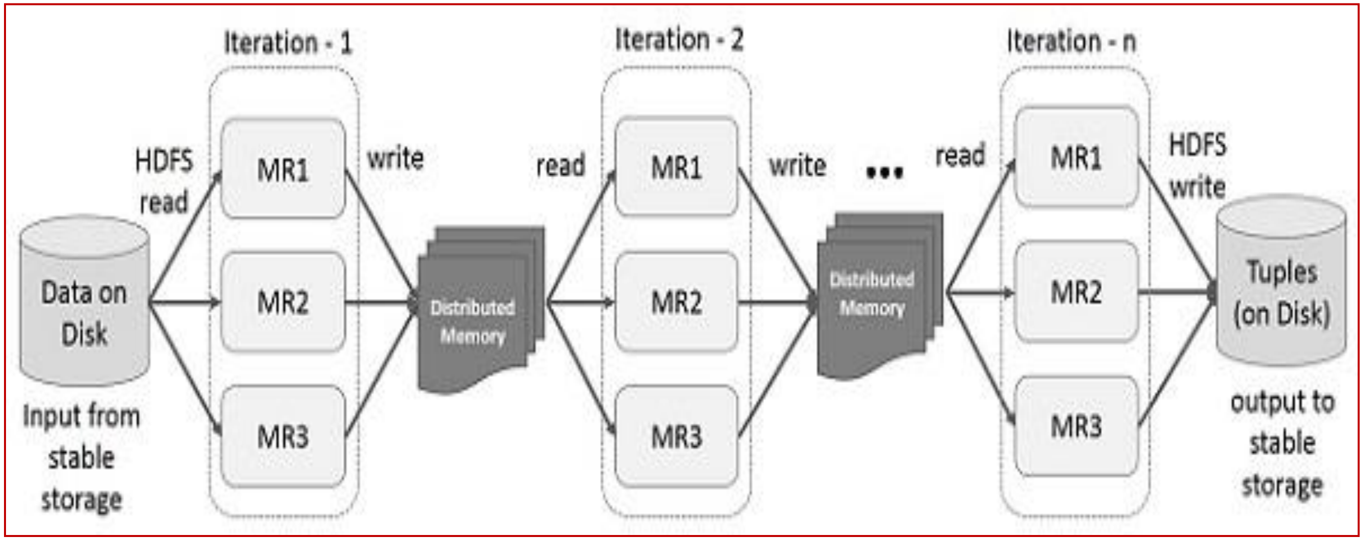
- Map Reduce is disk based

- If we can retrieve disk data into memory and then use it for further processing, we can get even better response time

Difference between Hadoop & Spark



Iterative operation using Hadoop using disk



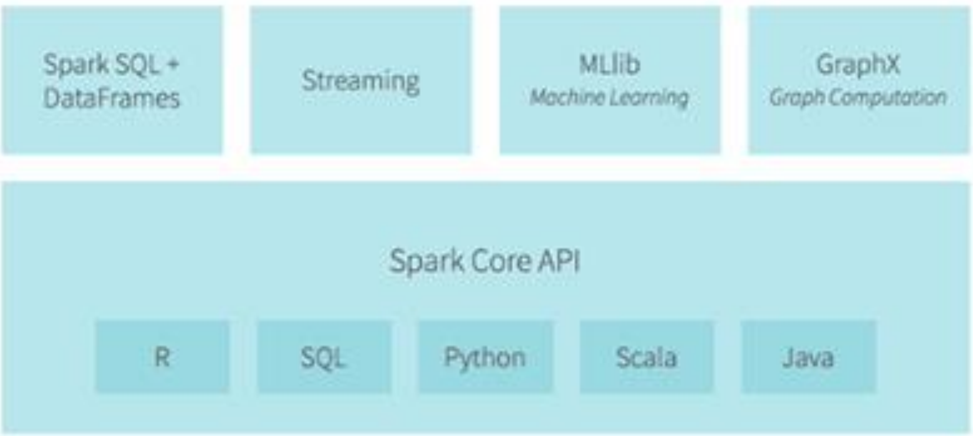
Iterative operation using Spark using memory

Source: https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm

Apache Spark

- Apache Spark is open source, general-purpose distributed computing engine used for processing and analyzing a large amount of data
- Main feature: In-memory cluster computing
- Useful for real time computations

Apache Spark Components



Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

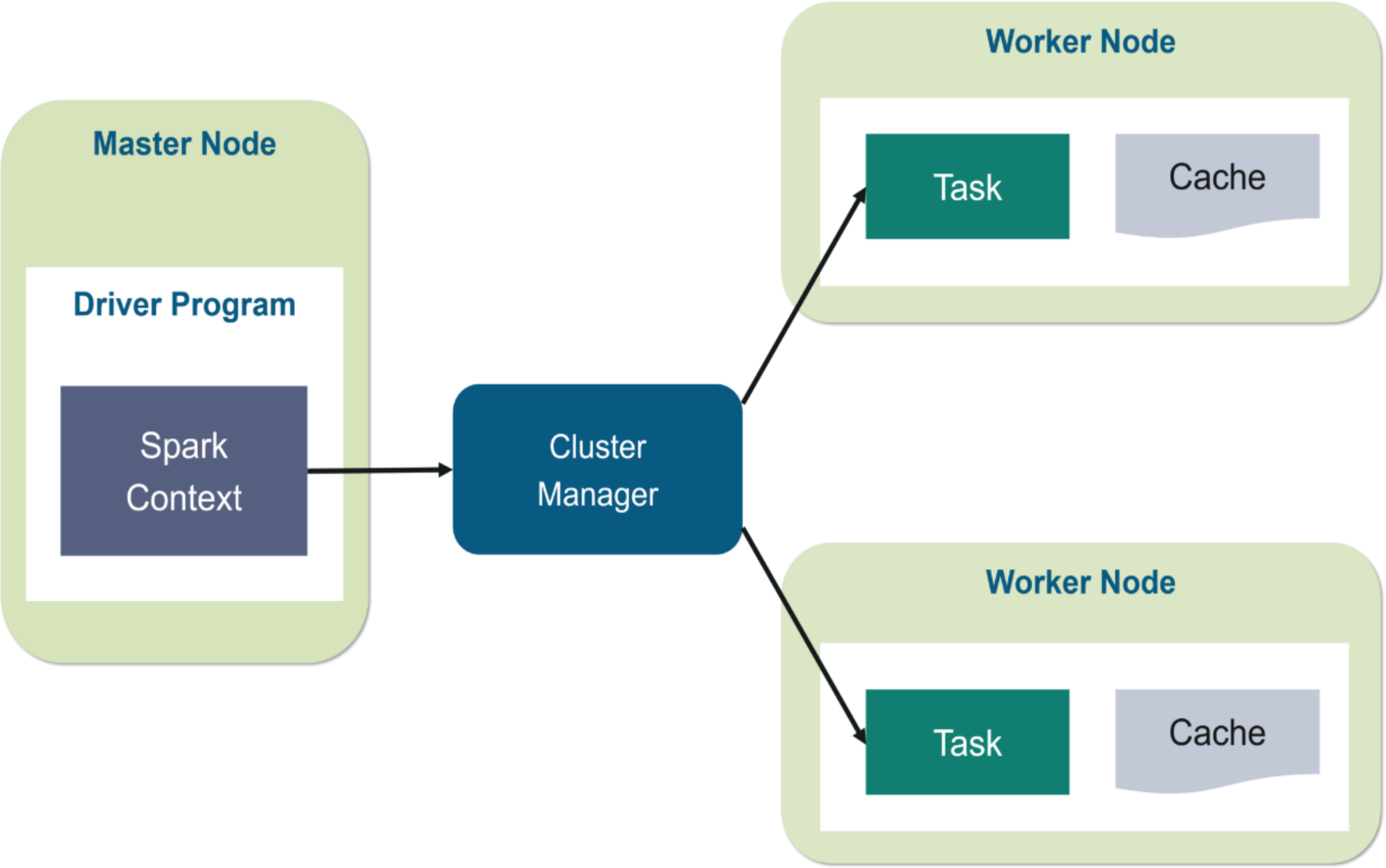
SQL

Python

Scala

Java

Spark architecture

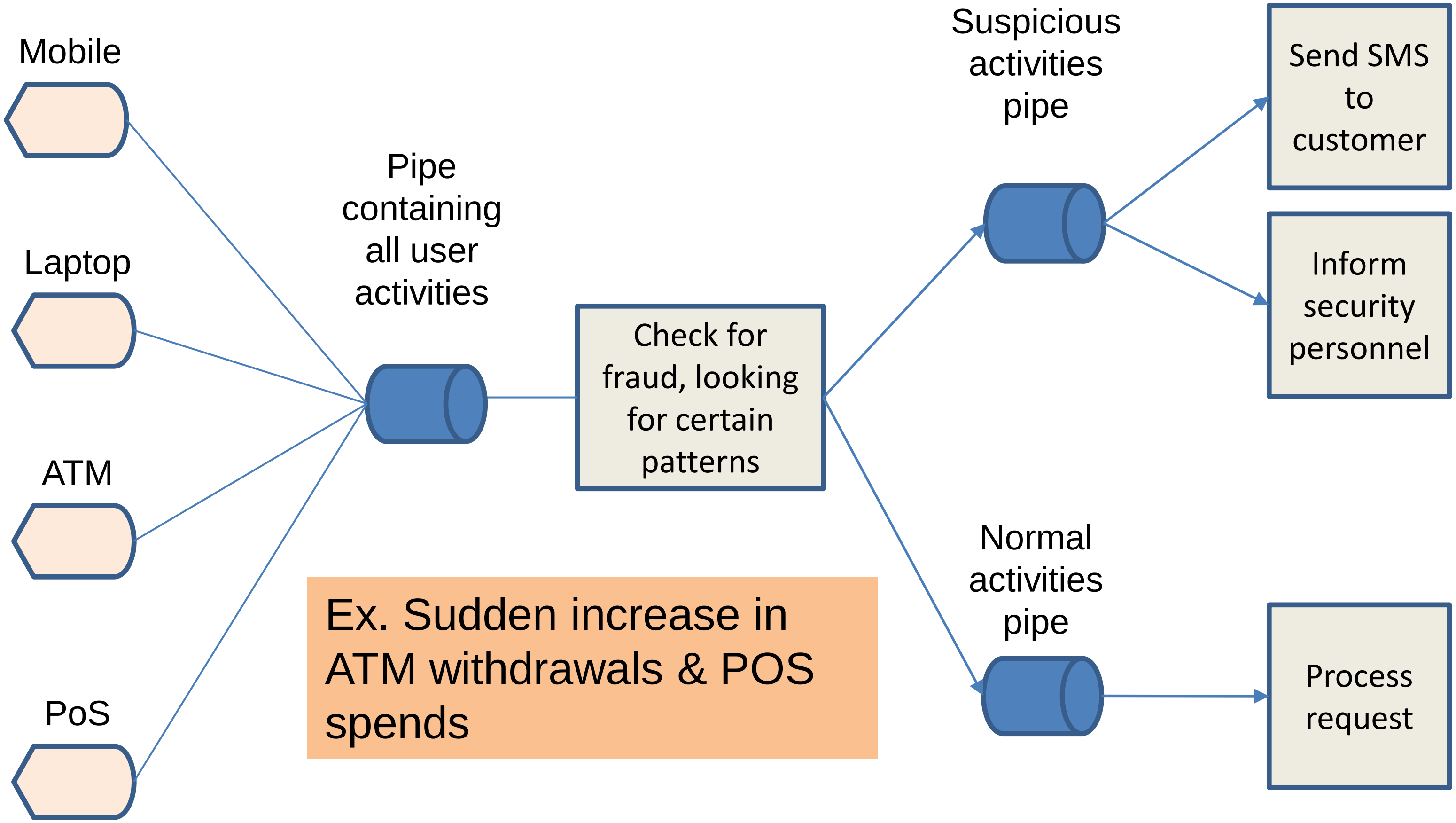


Real time analytics

- Real time analytics lets users see, analyze and understand data as it arrives in a system.
- It can give users insights for making real-time decisions.
- Examples:
  - Real time advertising
  - Identify security breaches
  - Sensor data processing to predict issues in machines

Real time analytics - Fraud detection in bank

Continuous monitoring of client's activity to see if there are any potential issues


Mobile
Laptop
ATM
PoS
Pipe containing all user activities
Check for fraud, looking for certain patterns
Suspicious activities pipe
Normal activities pipe
Send SMS to customer
Inform security personnel
Process request

Ex. Sudden increase in ATM withdrawals & POS spends

Real time analytics at Dream11 – a fantasy sports platform

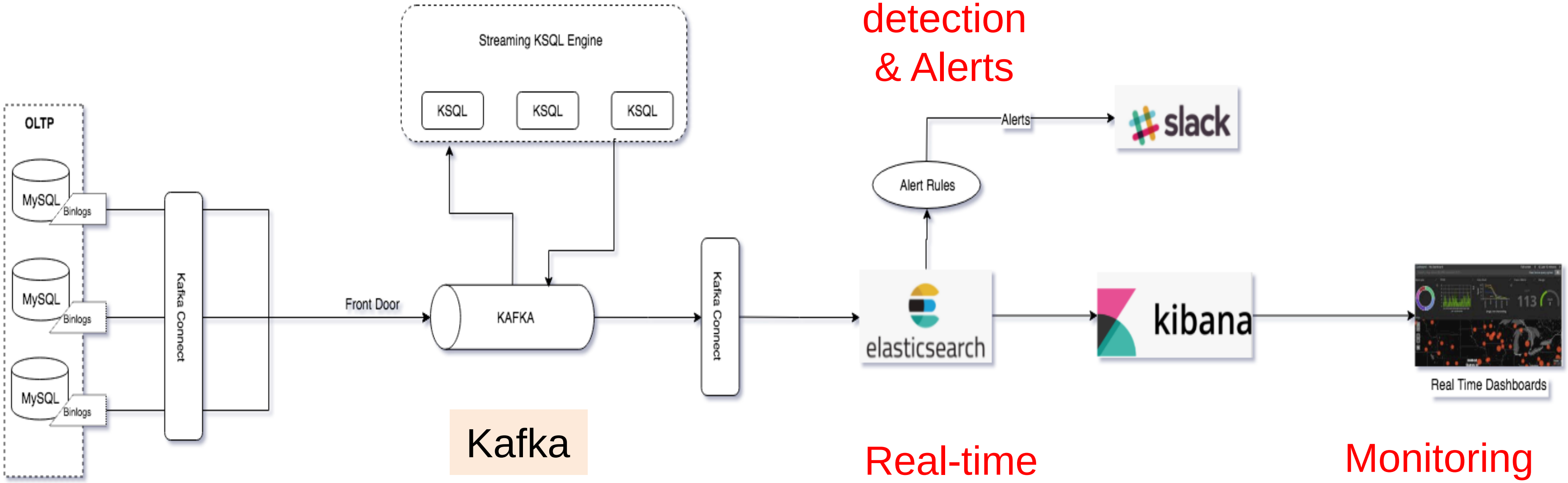
Objectives:

1. Know the real-time rate of contest joins
2. Know the real-time aggregated status of payment gateways
3. Identify real-time anomalies eg: unusual traffic on the system
4. Realtime aggregated view of outcome of marketing campaigns
5. How customers are using discount coupons once promotion goes live
6. Realtime alerting once Mega contest is above 90%

Real time analytics at Dream11 – a fantasy sports platform

Perform data enrichment

Anomaly detection & Alerts



Kafka

Transaction systems

Real-time data store

Monitoring

Dream11 case study

Ref: [medium.com]

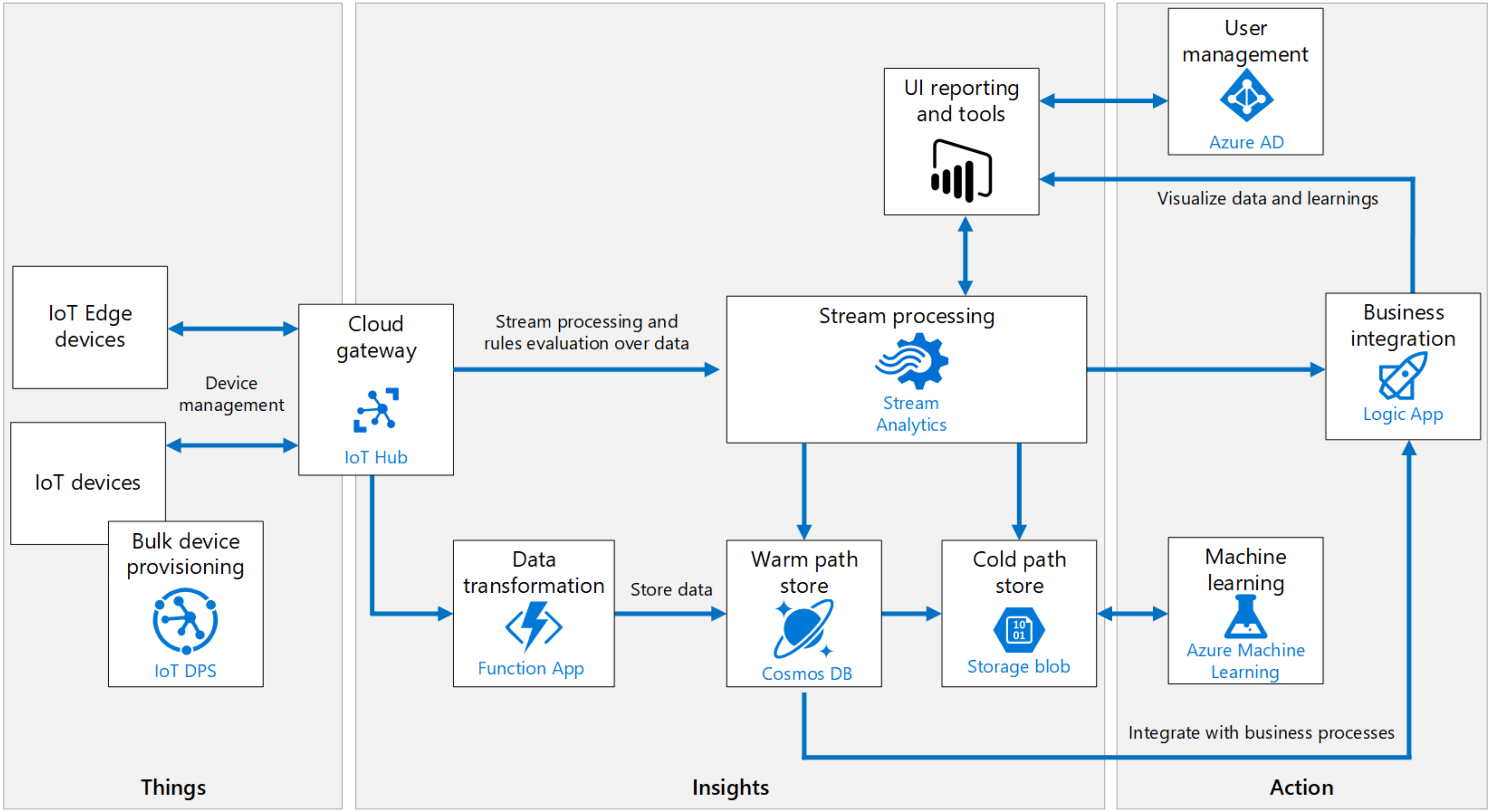
IoT: Internet of Things

Inter-connected devices which work together and perform operations with little human intervention

Examples:

- Tracking machine parameters using sensors and controlling for optimum performance
- Tracking goods, real time information exchange about inventory among suppliers and retailers
- Sensing for soil moisture and nutrients, controlling water usage for plant growth

Azure IoT reference architecture



Ref: docs.microsoft.com

Appendix