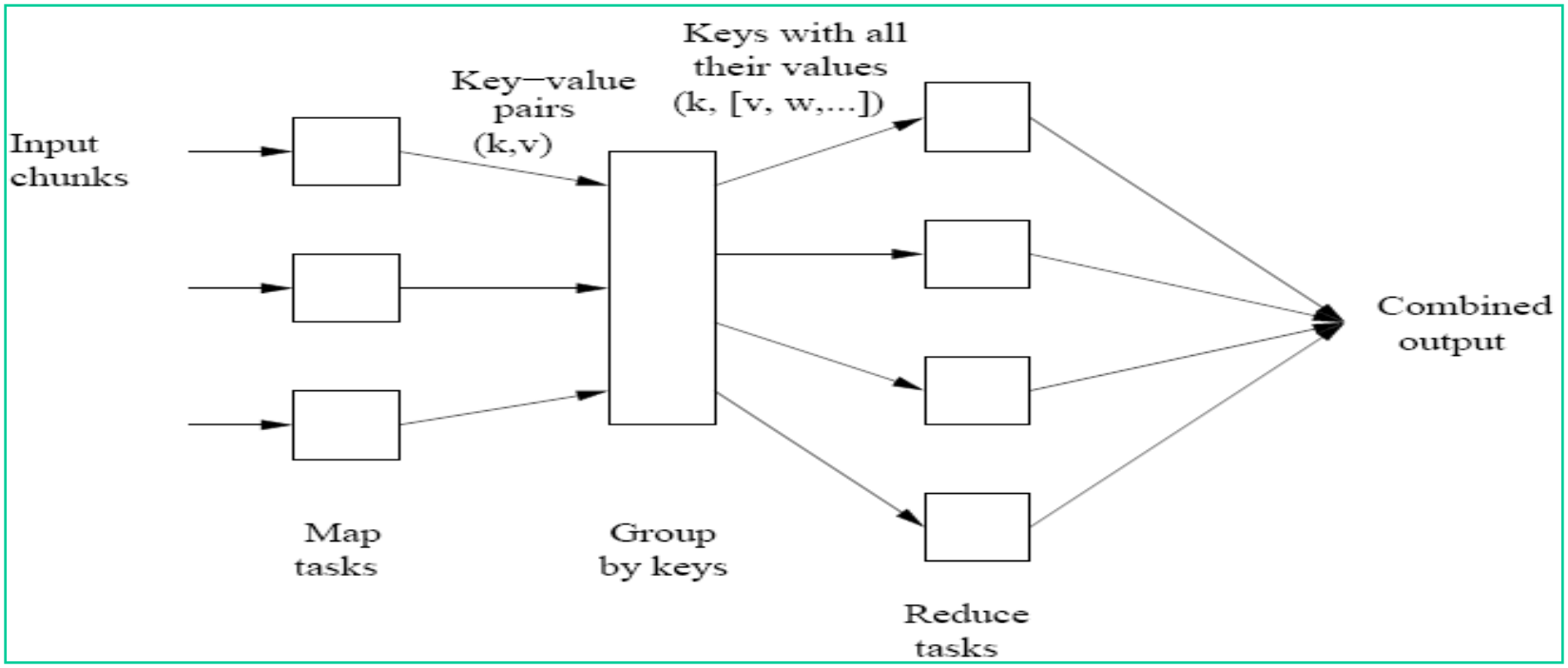
Big Data Processing Architecture

MAP REDUCE

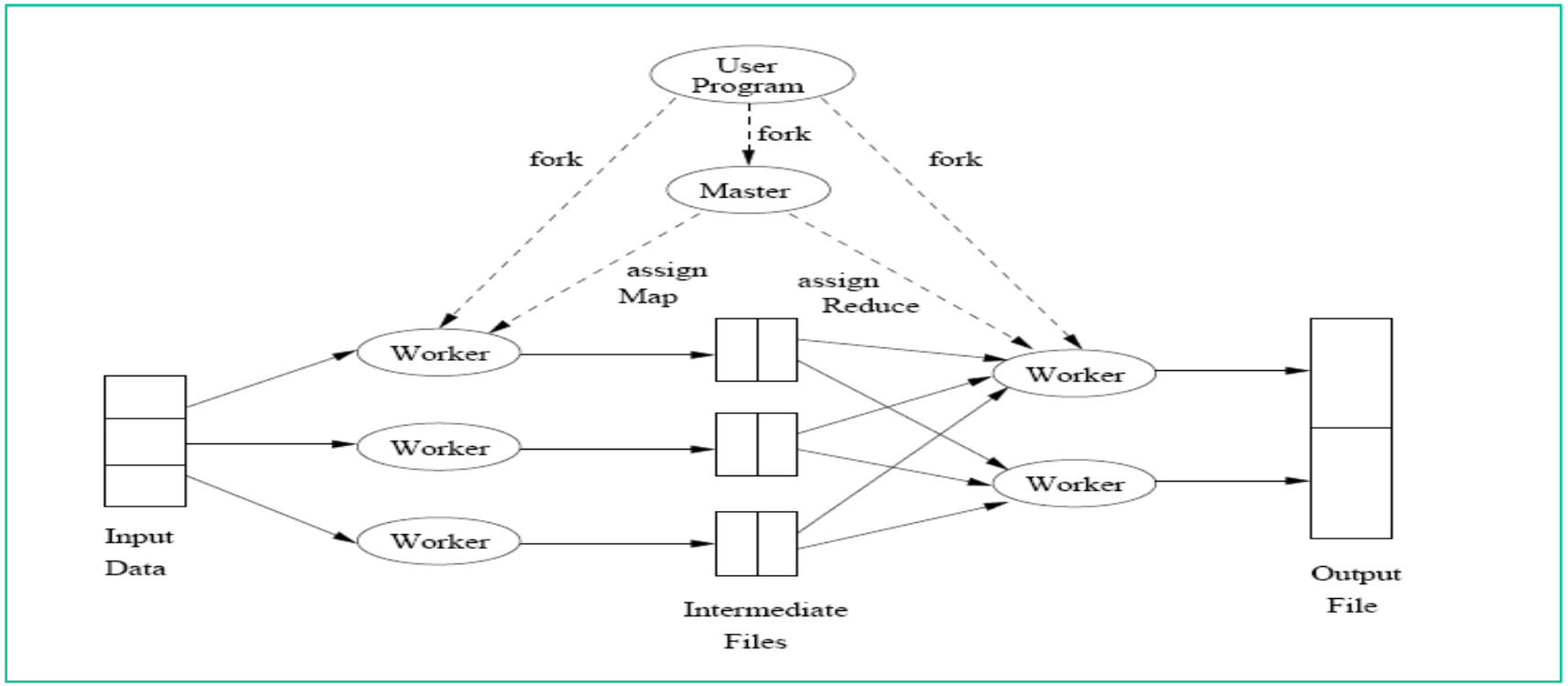
Map Reduce: Google's Invention

- Map: User program that processes input to generate (key, value) pairs
- Reduce: User programs that act on the data sorted on 'key' of Map to generate the output



Map Reduce: Physical Architecture

- Worker Node: Can run on commodity hardware
- Master Node: Normal server scale hardware
- Connectivity: Gigabit per second throughput essential



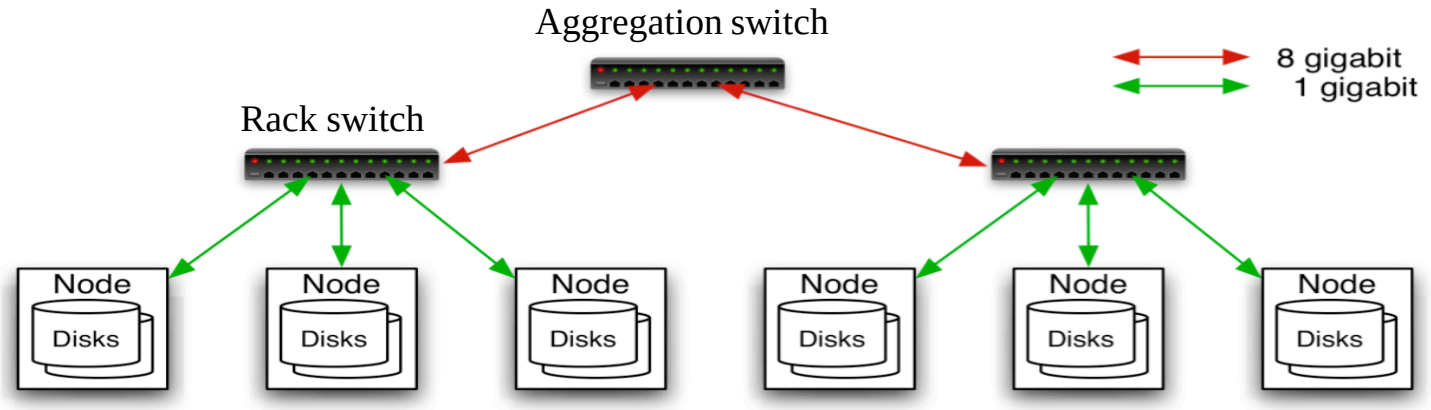
Hadoop - Why ?

- Need to process huge datasets on large clusters of computers
- Very expensive to build reliability into each application
- Nodes fail every day
  - Failure is expected, rather than exceptional
  - The number of nodes in a cluster is not constant
- Need a common infrastructure
  - Efficient, reliable, easy to use
  - Open Source, Apache Licence

Who uses Hadoop?

- Amazon/A9
- Facebook
- Google
- New York Times
- Yahoo!
- Oracle
- …. many more

Commodity Hardware



- Typically in 2 level architecture
  - Nodes are commodity PCs
  - 30-40 nodes/rack
  - Uplink from rack is 8 gigabit
  - Rack-internal is 1 gigabit

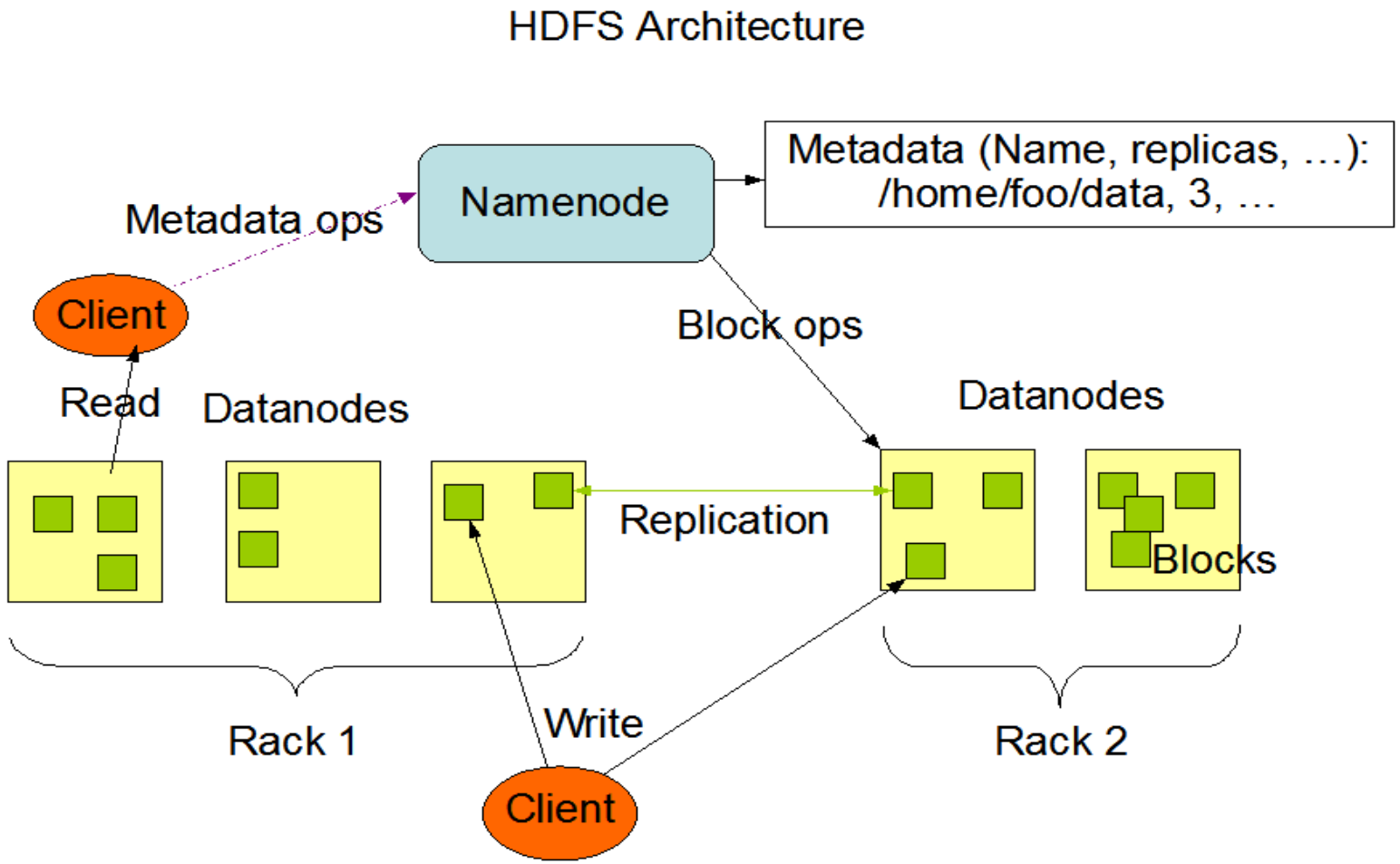
Goals of Hadoop Distributed File System

- Very Large Distributed File System
  - 10K nodes, 100 million files, 10PB
- Assumes Commodity Hardware
  - Files are replicated to handle hardware failure
  - Detect failures and recover from them
- Optimized for Batch Processing
  - Data locations exposed so that computations can move to where data resides
  - Provides very high aggregate bandwidth

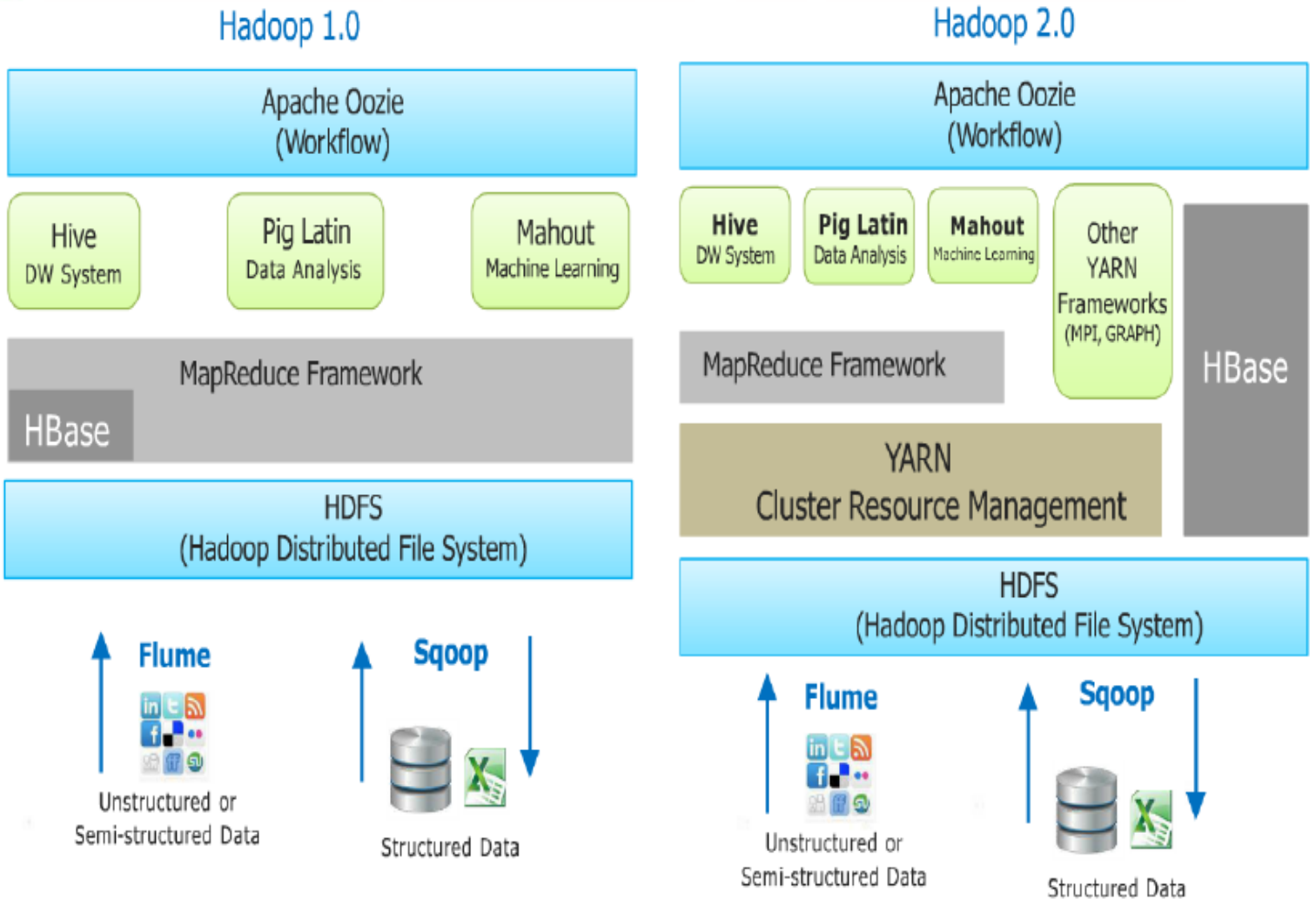
Hadoop Distributed File System

- Single Namespace for entire cluster
- Data Coherency
  - Write-once-read-many access model
  - Client can only append to existing files
- Files are broken up into blocks
  - Typically 64MB block size
  - Each block replicated on multiple DataNodes
- Intelligent Client
  - Client can find location of blocks
  - Client accesses data directly from DataNode

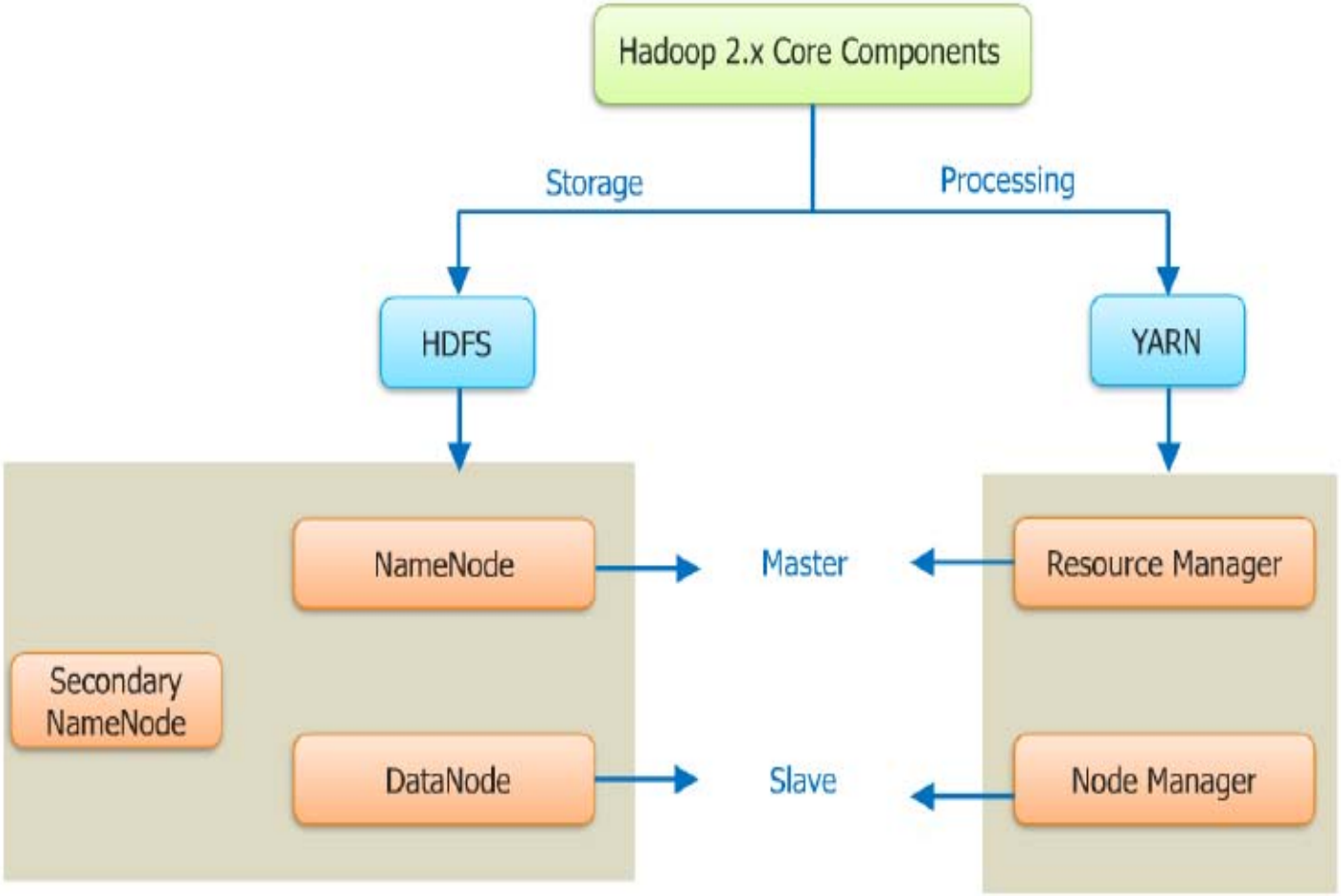
HDFS Architecture



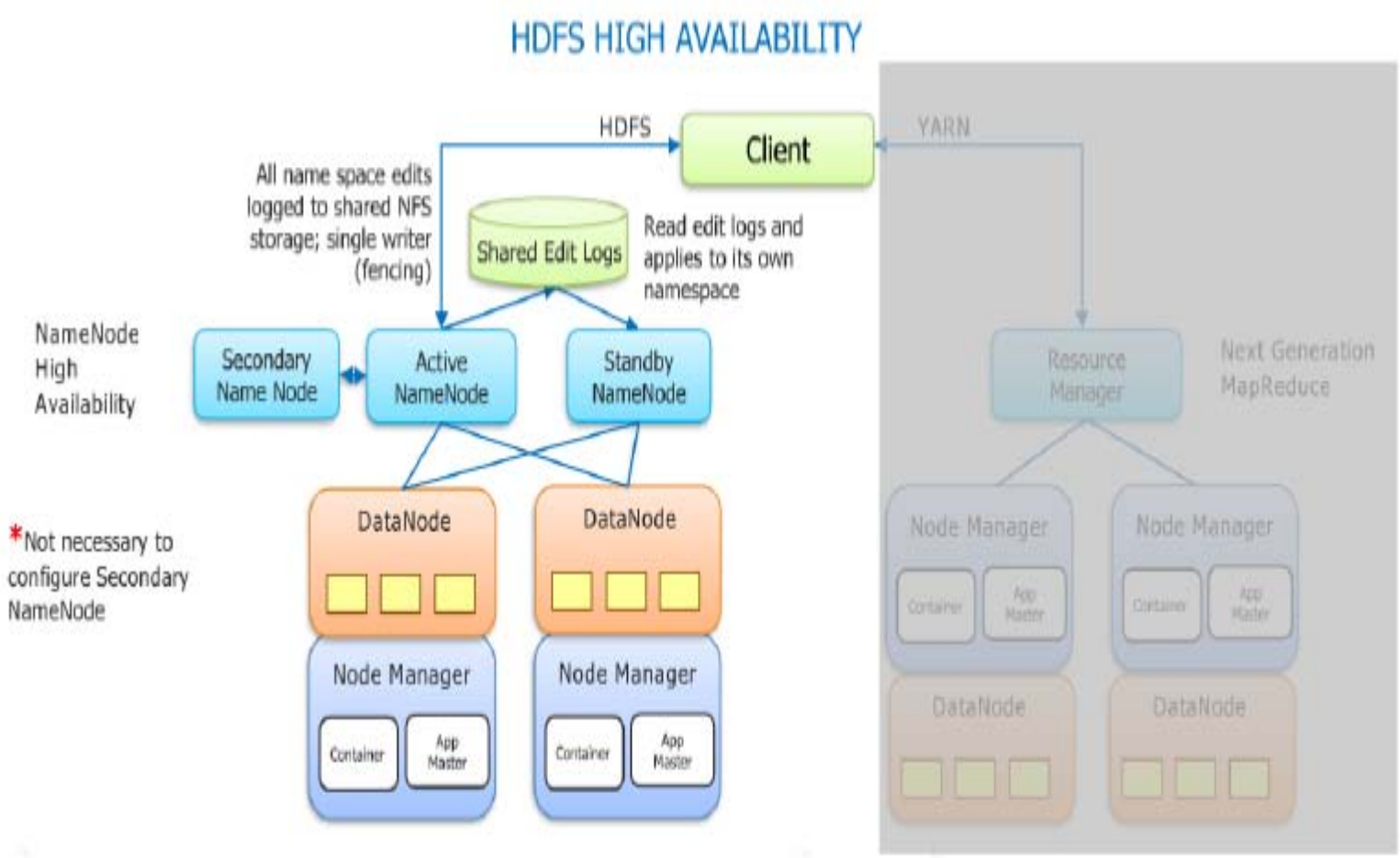
Hadoop Ecosystem



Hadoop 2.x Core Components

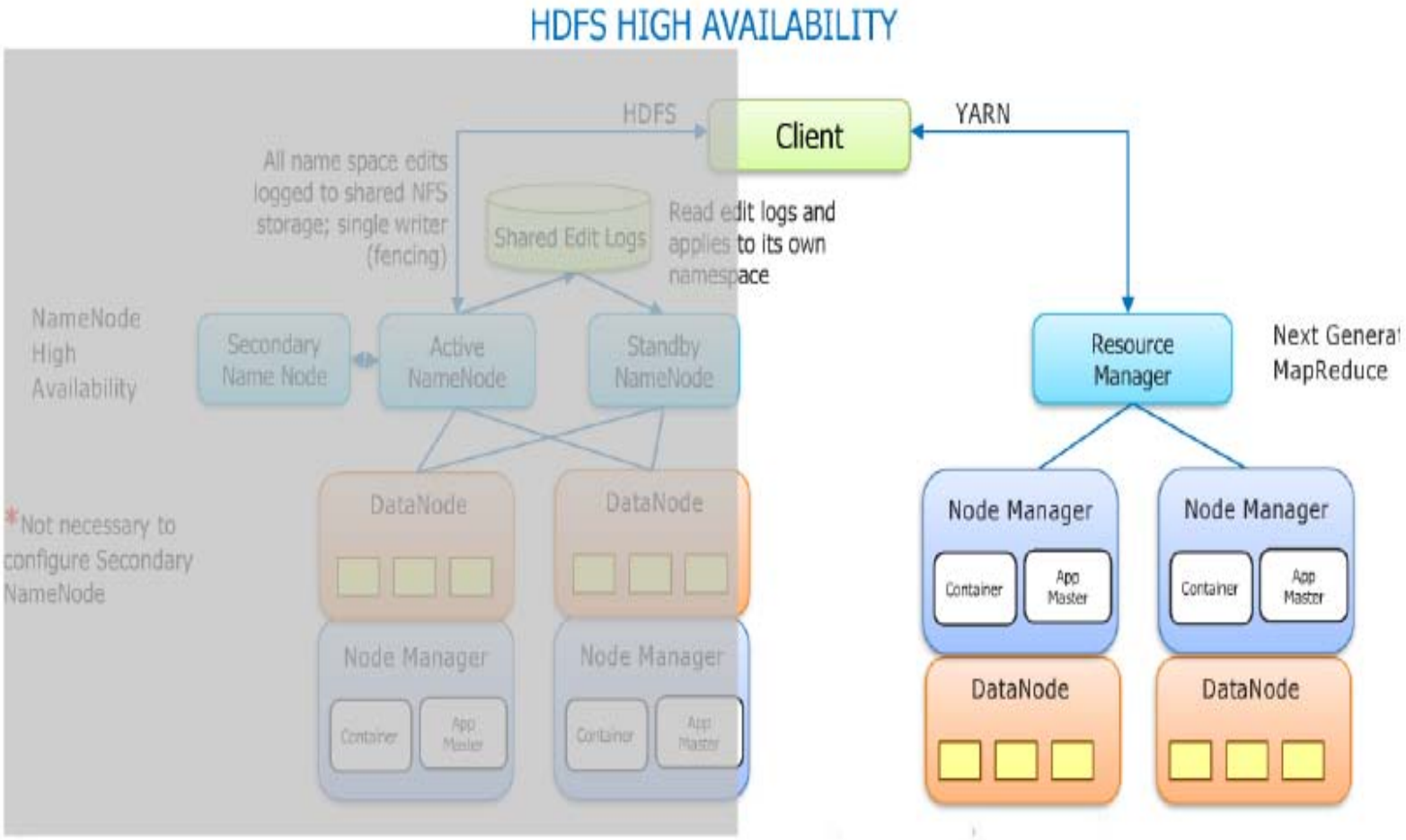


Hadoop 2.x High Availability



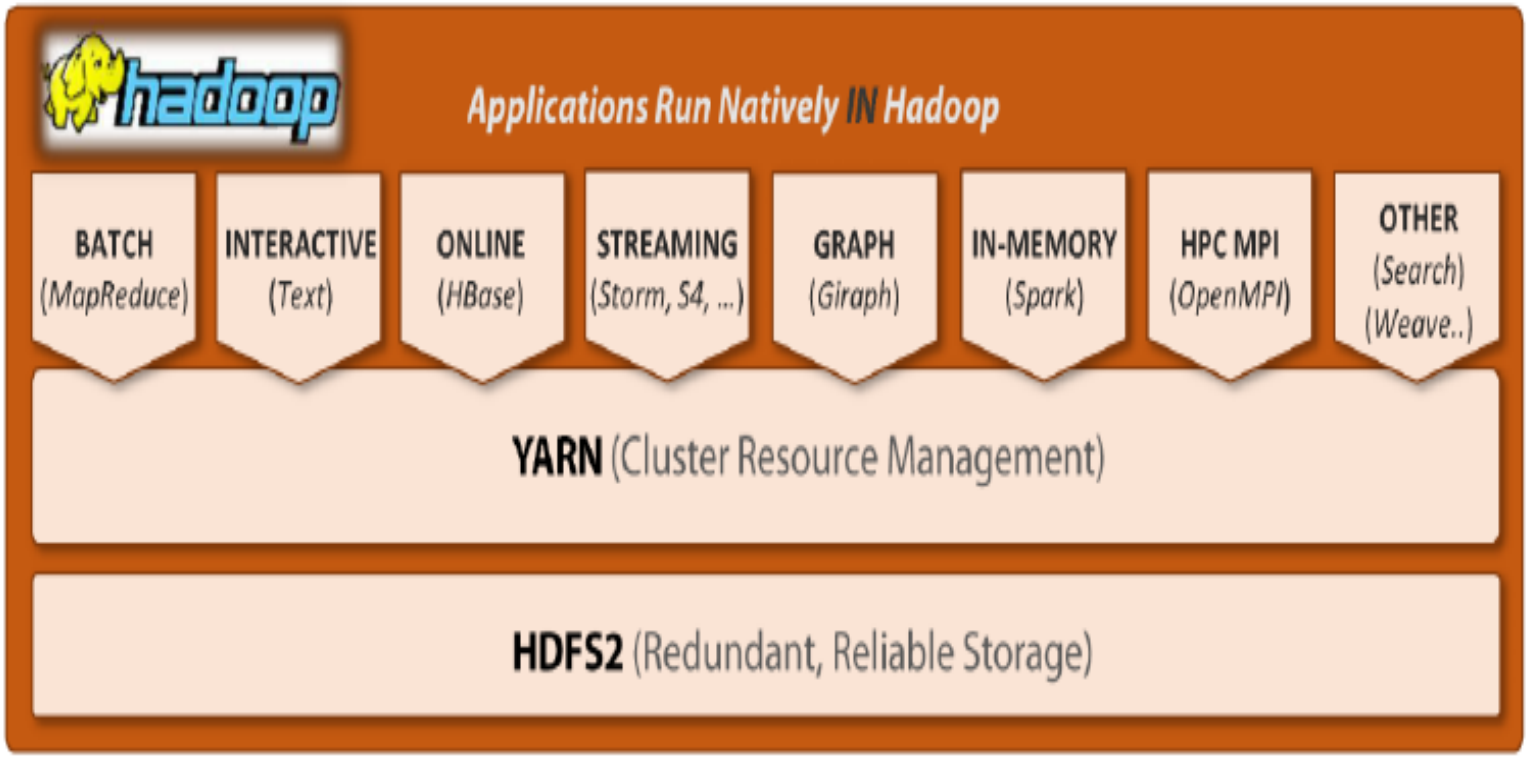
http://hadoop.apache.org/docs/stable2/hadoop-yarn/hadoop-yarn-site/HDFSHighAvailabilityWithNFS.html

Hadoop 2.x High Availability



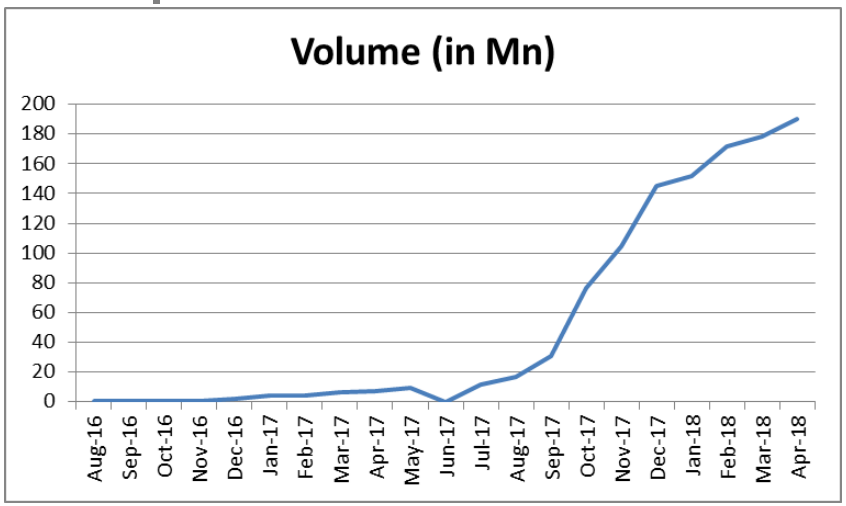
http://hadoop.apache.org/docs/stable2/hadoop-yarn/hadoop-yarn-site/HDFSHighAvailabilityWithNFS.html

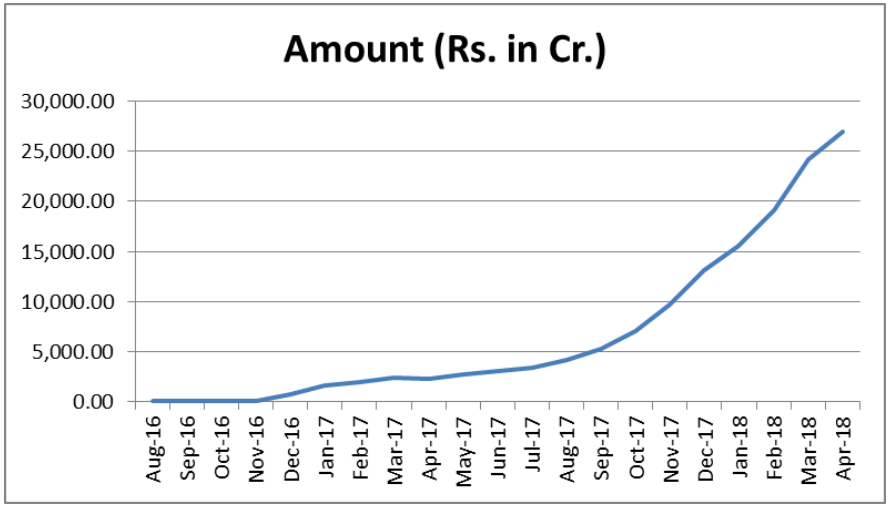
YARN: Beyond Map-Reduce



http://hadoop.apache.org/docs/stable2/hadoop-yarn/hadoop-yarn-site/YARN.html

UPI Transactions





Other than Facebook no application worldwide has seen such exponential adoption in early years of launch

Business Questions : Customer Focus

- Google : How many new users have logged since last month

- Uber: What should be the dynamic price that would not put off customer

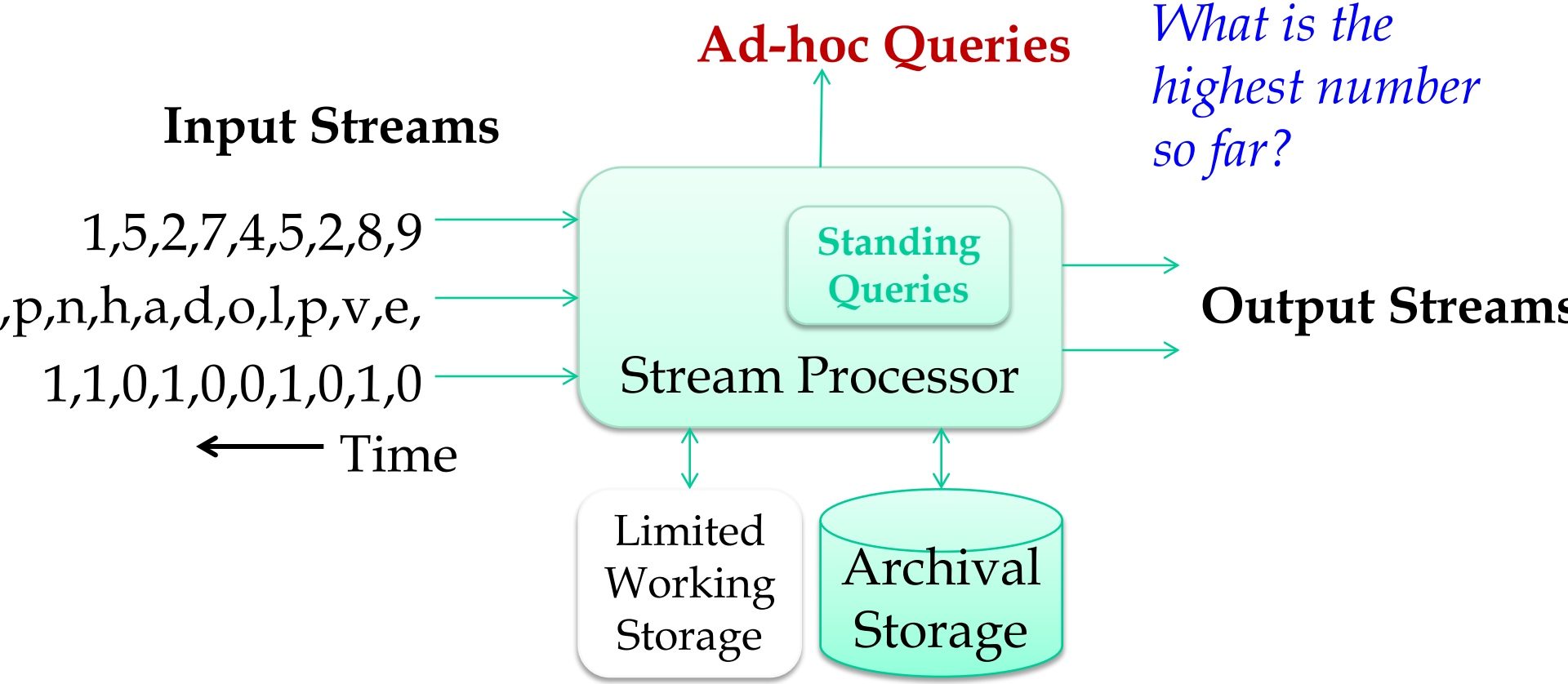
Considerations and Approach

- Constraint
  - We cannot store all the data from the past
  - The stream objects can only be scanned once
  - Usually we need to answer it in seconds as it provides an opportunity for the business
- General Approach
  - Obtain a close approximate answer within a short time

Stream Model

How many new logins in past week?

What is the highest number so far?



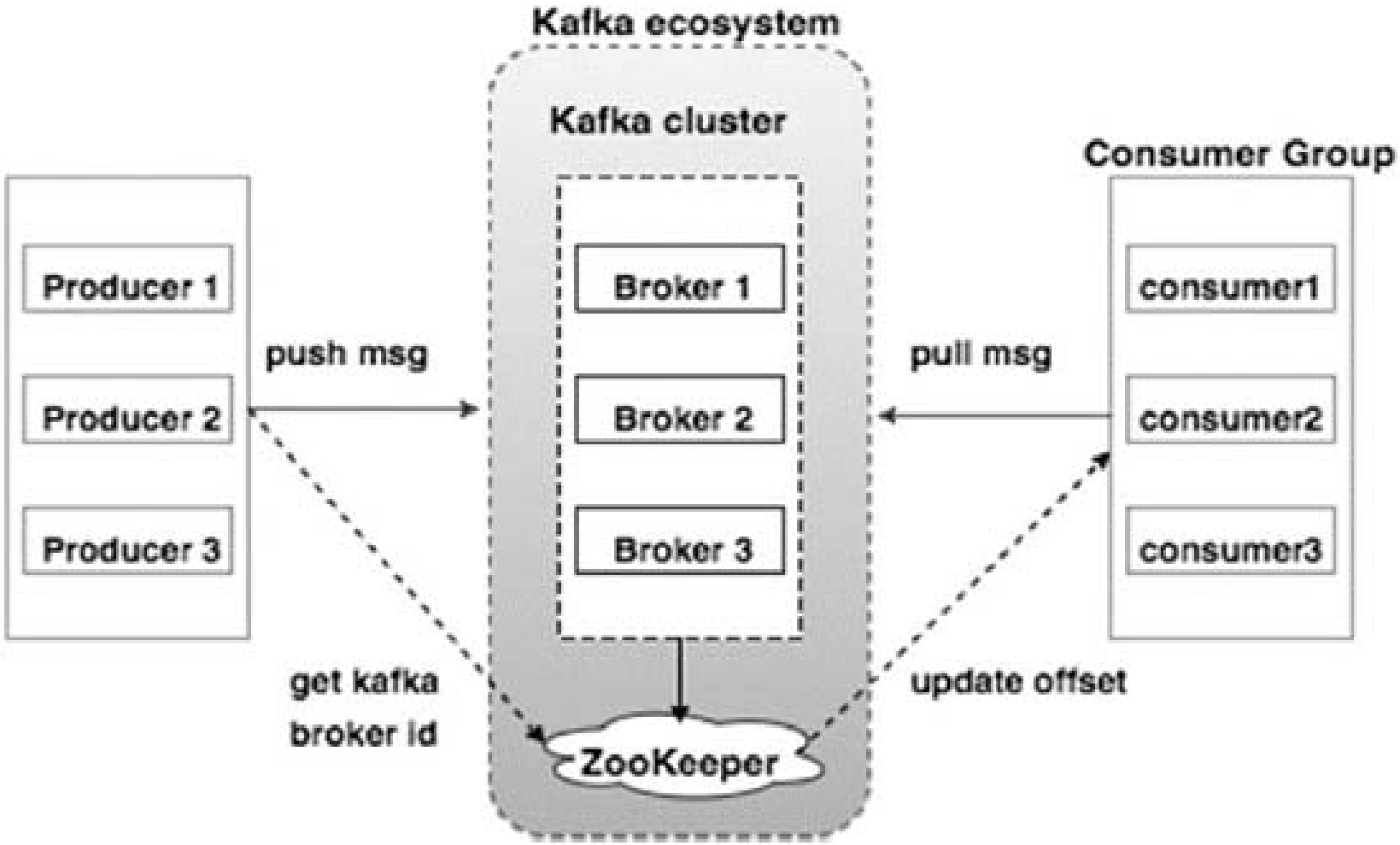
Data Streams: Mining data from the flow

- Challenges / Techniques
  - **Sampling**, without loss of characteristics
  - **Filtering**, selecting the elements that **belong to a set** and discarding the rest
  - **Distinct Elements**, using statistical functions to arrive at counts of distinct elements
  - **Standing Queries**, to "collect" the answers in the fly
  - **Decaying Time Windows**, to weight the properties in the past as a weight of time

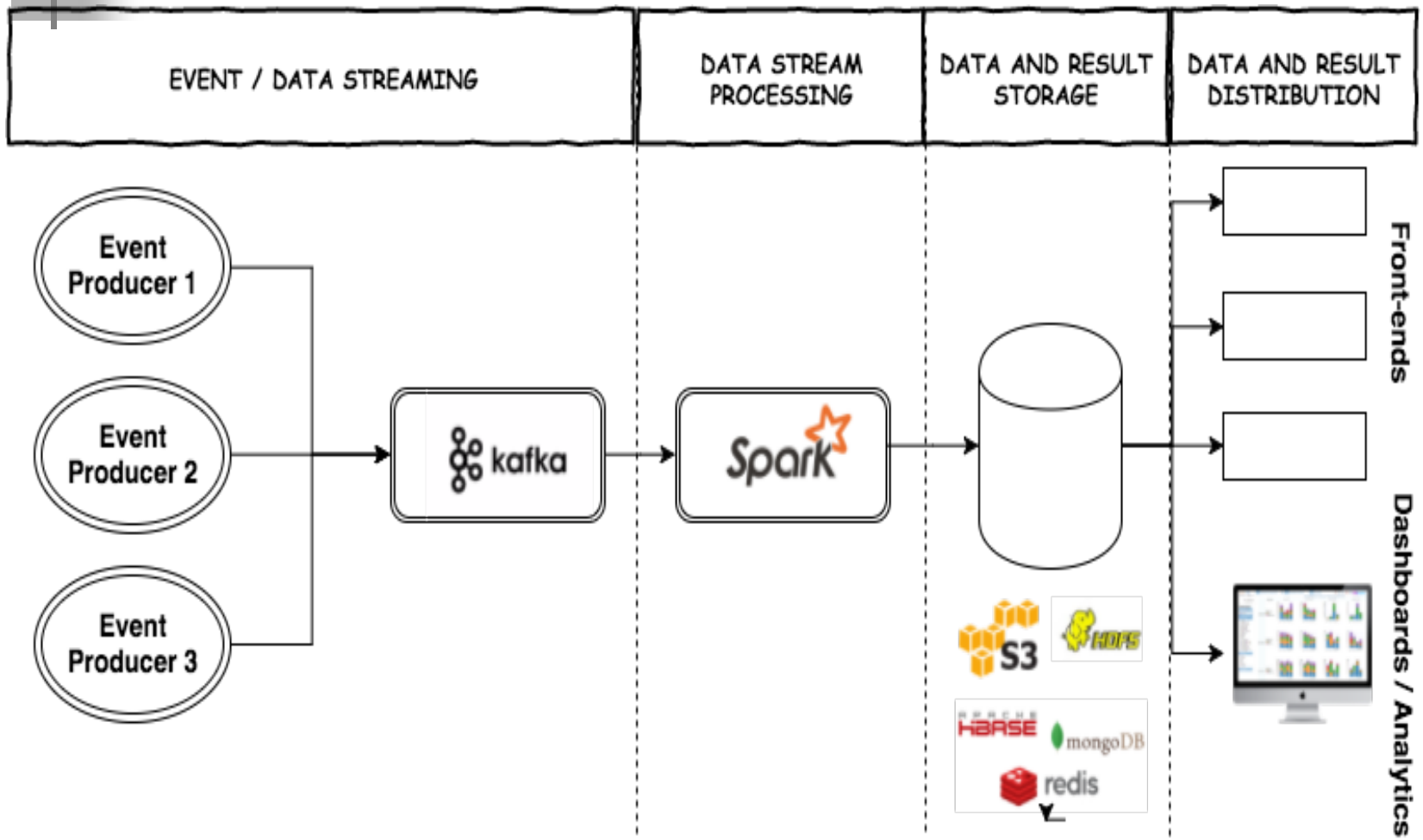
Stream Processing needs

- Messaging Software on Clusters along with Coordination system
  - Typically combination of Apache Kafka (Publish-Subscribe Messaging) and Apache Zookeeper
- Stream Processing software
  - Typically Apache Storm or Spark
- In-memory Storage and Database
  - Typically Redis

Kafka Architecture



Kafka – Spark Integration



Data Ingestion

DATA INGESTION

Def: Data ingestion is a process by which data is moved from one or more sources to a destination where it can be stored and further analysed.

- Data from different sources and in different formats (RDBMS, other types of databases, Text files, S3 buckets (Amazon), CSVs, or from streams)
- Needs to be cleansed and transformed in a way that allows us to analyse it together with data from other sources.
- Can be done in real time, in batches, or in a combination of the two (Lambda architecture)

Ref:www.alooma.com

Why do we need Data Ingestion:

- **Speed**: Data has gotten to be much larger, more complex and diverse, and the old methods of data ingestion just aren't fast enough.

- **Complexity:** Due to variety and velocity at which data gets generated, cleansing becomes more complex.

- **Centralizing** data makes it easy for anyone at the company (or outside of the company, depending on goals) to retrieve, inspect, and analyse it.

Data Ingestion tools

- Apache kafka
- Apache Nifi
- Wavefront
- DataTorrent
- Amazon Kinesis
- Syncsort
- Gobblin
- Apache Flume

etc….

Tools(Contd.)

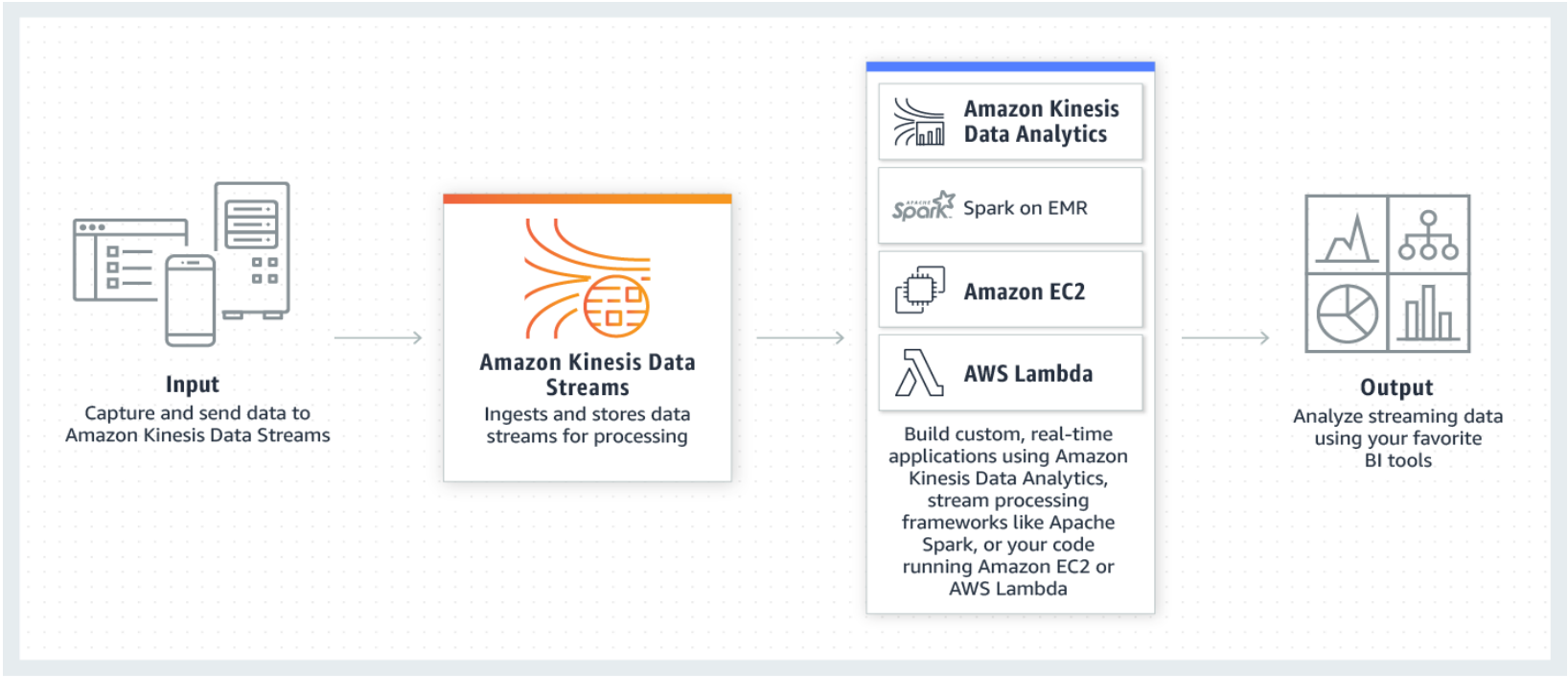
- **Apache Kafka:** Apache Kafka is a distributed publish-subscribe messaging system and a robust queue that can handle a high volume of data and enables you to pass messages from one end-point to another. Kafka is suitable for both offline and online message consumption.

- **Apache Nifi:** Apache NiFi is an open source project which enables the automation of data flow between systems, known as "data logistics". The project is written using flow-based programming and provides a web-based user interface to manage data flows in real time.

- **Gobblin:** A distributed data integration framework that simplifies common aspects of big data integration such as data ingestion, replication, organization and lifecycle management for both **streaming** and **batch** data ecosystems.

Amazon Kinesis

- With Amazon Kinesis, real-time data can be ingested such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications. Amazon Kinesis enables user to process and analyze data as it arrives and respond instantly instead of having to wait until all data is collected before the processing can begin.



Benefits of Amazon Kinesis

- **Real-time:** Amazon Kinesis enables user to ingest, buffer, and process streaming data in real-time, so you can derive insights in seconds or minutes instead of hours or days.

- **Fully managed:** Amazon Kinesis is fully managed and runs the streaming applications without requiring user to manage any infrastructure.

- **Scalable:** Amazon Kinesis can handle any amount of streaming data and process data from hundreds of thousands of sources with very low latencies.

Data Formats

- Text / CSV
- JSON (used when data is sent from a server to a web page)
- Sequence File (binary key-value pair format)
- Avro (Language neutral data serialization system)
- Parquet (stores nested data in columnar format using definition and repetition levels)
- ORC (Optimized row columnar format)
- XML

Data Discovery

Def: Data discovery is the process of obtaining actionable information by finding patterns in data from multiple sources with interactive visual analysis.

It can be clustered into three main categories:

- data preparation.
- visual analysis.
- guided advanced analytics.

Data Discovery- Contd.

- Requires skills on understanding data relationships and data modelling as well as using data analysis and guided advanced data analytics functions
- Data preparation helps business users to connect to relevant enterprise and external data sources.
- Guided advanced analytics functions provide statistical information on data which users can employ for more sophisticated and pattern oriented data analysis.

Tools:

- Must be **simple to implement , adaptable, code free environment for business users**(no need for statistical degrees and an analytical background to use it.)
- Can easily work with massive amounts of data.
- Must be fast so that decision making is done quickly.
- Must provide simple-to-understand depictions of data via visualizations.

Data Acquisition

- **Data acquisition** has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.

- The acquisition of big data is most commonly governed by four of the Vs: **volume, velocity, variety, and value**

- Most data acquisition scenarios assume high-volume, high-velocity, high-variety, but low-value data, making it important to have adaptable and time-efficient gathering, filtering, and cleaning algorithms that ensure that only the high-value fragments of the data are actually processed by the data-warehouse analysis.

Big Data Value Chain

Data Acquisition	Data Analysis	Data Curation	Data Storage	Data Usage
• Structured data • Unstructured data • Event processing • Sensor networks • Protocols • Real-time • Data streams • Multimodality	• Stream mining • Semantic analysis • Machine learning • Information extraction • Linked Data • Data discovery • 'Whole world' semantics • Ecosystems • Community data analysis • Cross-sectorial data analysis	• Data Quality • Trust / Provenance • Annotation • Data validation • Human-Data Interaction • Top-down/Bottom-up • Community / Crowd • Human Computation • Curation at scale • Incentivisation • Automation • Interoperability	• In-Memory DBs • NoSQL DBs • NewSQL DBs • Cloud storage • Query Interfaces • Scalability and Performance • Data Models • Consistency, Availability, Partition-tolerance • Security and Privacy • Standardization	• Decision support • Prediction • In-use analytics • Simulation • Exploration • Visualisation • Modeling • Control • Domain-specific usage

ref: link.springer.com

Data Acquisition(Contd.)

To gather data from distributed information sources with the aim of storing them in scalable, big data-capable data storage, three main components are required:

- Protocols that allow the gathering of information for distributed data sources of any type (unstructured, semi-structured, structured)
- Frameworks with which the data is collected from the distributed sources by using different protocols.
- Technologies that allow the persistent storage of the data retrieved by the frameworks

Tools:

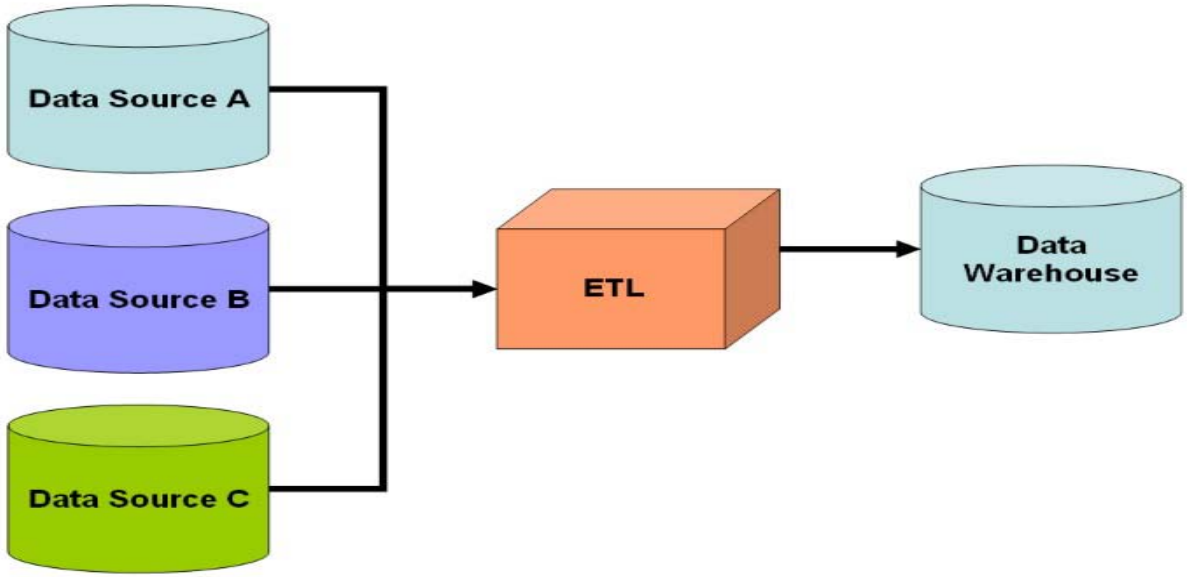
- **Storm-** It can be utilized in many data gathering scenarios including stream processing and distributed RPC for solving computationally intensive functions on-the-fly, and continuous computation applications

- **Kafka-** Kafka aims to unify offline and online processing by providing a mechanism for a parallel load into Hadoop as well as the ability to partition real-time consumption over a cluster of machines.

- **Flume-** purpose of Flume is to provide a distributed, reliable, and available system for efficiently collecting, aggregating, and moving large amounts of log data from many different sources to a centralized data store

Data Integration

Def: Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information. A complete data integration solution delivers trusted data from various sources. ETL – Extract-Transform-Load



Data Integration techniques

ETL is not anymore the only way to achieve the goal and that is a new level of complexity in the field of Data Integration.

There are many sophisticated ways the unified view of data can be created today.

1. Manual Data Integration

In this approach, a web based user interface or an application is created for users of the system to show them all the relevant information by accessing all the source systems directly. There is no unification of data in reality.

Data Integration techniques(Contd.)

2. **Middleware Data Integration** can act like a glue that holds together multiple legacy applications, making seamless connectivity possible without requiring the two applications to communicate directly.

3. **Data Virtualization Integration Approach**

- allows user to leave data in the source systems while allowing to create a new set of unified views.
- main benefit of the virtual integration approach is near real time view of data from the source systems.
- eliminates a need for separate data store for the consolidated unified data.

Data Integration techniques(Contd.)

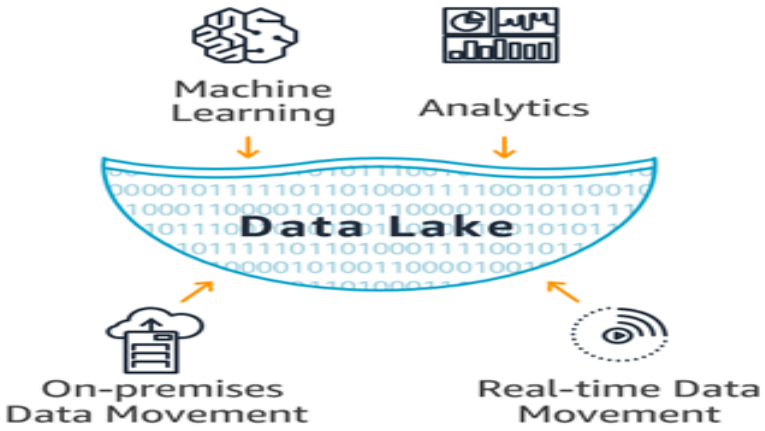
4. Data Warehouse Approach Of Data Integration

- requires creation of a new Data Warehouse (or Data Marts) which stores a unified version of data extracted from all the source systems involved and manage it independent of the original source systems.

- The benefits of this approach include ability to easily manage history of data (or data versioning), ability to combine data from very disparate sources (mainframes, databases, flat files, etc.) and to store them in a central repository of data.

What is Data Lake

Def: A data lake is a centralized repository that allows to store all structured and unstructured data at any scale. Data can be stored as-is, without having to first structure the data, run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.



Data Lakes compared to Data Warehouses

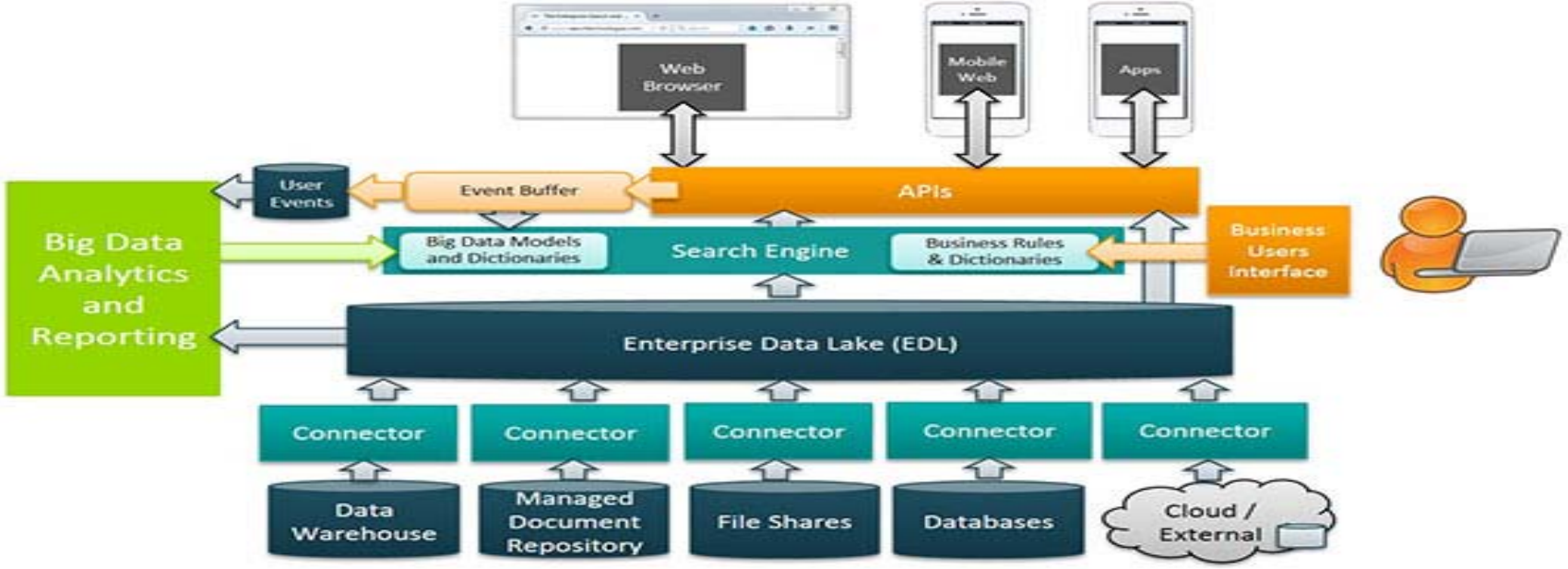
- A **data warehouse** is a database optimized to analyse relational data coming from transactional systems and line of business applications. The data structure, and schema are **defined in advance** to optimize for fast SQL queries, where the results are typically used for operational reporting and analysis. Data is cleaned, enriched, and transformed so it can act as the "single source of truth" that users can trust.

- A **data lake** is different, because it stores relational data from line of business applications, and non-relational data from mobile apps, IoT devices, and social media. The structure of the data or **schema is not defined** when data is captured. Different types of analytics on the data like SQL queries, big data analytics, full text search, real-time analytics, and machine learning can be used to uncover insights.

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

A Data Lake Architecture

All data will be ingested into the data lake or staging repository (based on Cloudera) and then searched (using a search engine such as Cloudera Search or Elasticsearch). Where necessary, content will be analysed and results will be fed back to users via search to a multitude of UIs across various platforms.



Data Fusion

Def: **Data fusion** is the process of integrating multiple data sources to produce more consistent, accurate, and useful information than that provided by any individual data source.

- Often categorized as low, intermediate, or high, depending on the processing stage at which fusion takes place.

- Low-level data fusion combines several sources of raw data to produce new raw data. The expectation is that fused data is more informative and synthetic than the original inputs.

Components of Data Fusion

- DB/search engine selector-
  Selects system to use

- Query dispatcher-
  Submit queries to selected search engines

- Document selector-
  Select document to use

- Result merger
  Merge selected document results

Bias Concept

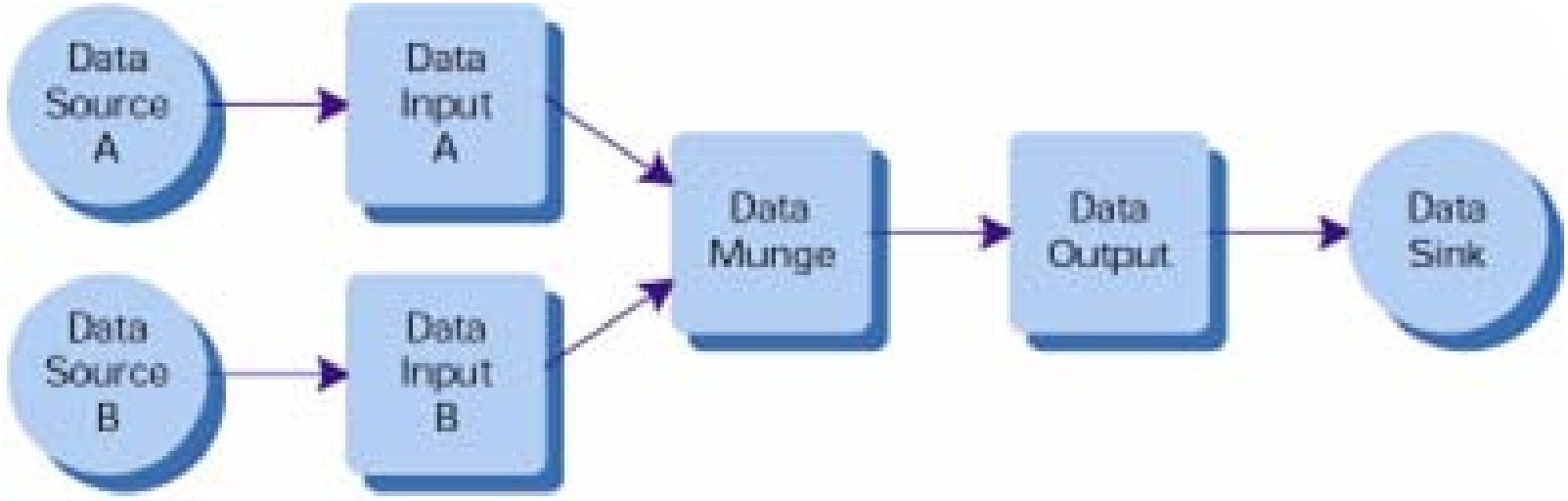
- Certain percentage of systems that behave differently from the norm(majority of the systems) are used.

- Biased systems improve data fusion
  - ➢ Eliminates ordinary systems from the fusion
  - ➢ Better discrimination between documents and systems

Data Munging

DATA MUNGING

Def: Data munging, also known as data wrangling is the process of converting and mapping data from its raw form to another format with the purpose of making it more valuable and appropriate for advance tasks such as Data Analytics and Machine Learning.



Goals of data munging:

- Provide precise and actionable data to Business analysts in a scheduled manner.
- Reduce time spent on collecting and arranging data
- Allow Data Scientist to focus more on analysis of data rather than wrangling of data
- Drive better decisions based on data in short time span

Steps in data munging:

1. Acquiring data: involves acquiring and sorting of data
2. Data cleaning: tedious but fundamental step of data munging process. Python can be used to clean up data.

Data Munging Tools

- **"R" packages**
  R is an important programming language for data scientists. It supports statistical and probability functions, and excels at handling slabs of numeric data.
- **Data Wrangler**
  DataWrangler is an interactive tool for data cleaning. It takes messy, real-world data and transforms it into data tables which can be exported to Excel, Tableau, R, etc.
- **Python and Pandas**
  Python and its Pandas library, which includes its DataFrame object, helps data scientists perform complex operations. For example, merging, joining, and transforming huge chunks of data with a single Python statement.
- **CSVKit**
  Helps convert data – from Excel to CSV, JSON to CSV.
- **Mr. Data Converter**
  Takes Excel data and transforms it to web-friendly formats like HTML, JSON and XML.

Data Reduction:

Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost same) analytical result.

Why?

- A database/ data warehouse may store terabytes of data.
- Complex data analysis/ mining may take a very long time to run on the complete data set.
- Impractical or infeasible analysis.

Dimensionality Reduction

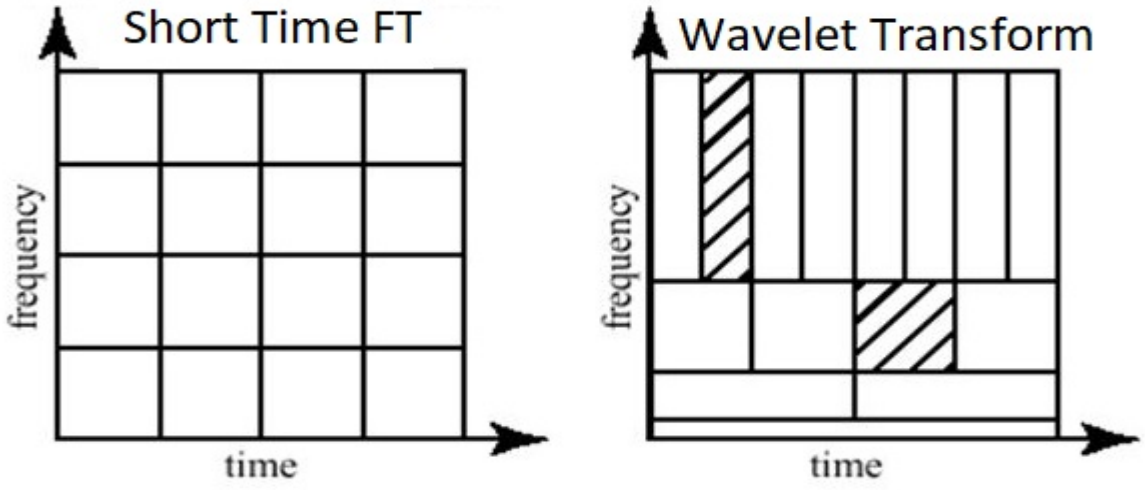
It is the process of reducing the number of random variables or attributes under consideration.

Dimensionality reduction methods include:

- Wavelet transforms
- Principal Component Analysis
- Attribute subset selection

Wavelet Transforms

- An approach for analyzing signals with a dynamical frequency spectrum is the Wavelet Transform.
- Has a high resolution in both the frequency- and the time-domain.
- Not only tells which frequencies are present in a signal, but also tells the time of their occurrence.



Principal Component Analysis

Technique to emphasize variation and bring out strong patterns in a dataset. Can be performed in R.

Given N data vectors from n-dimenions, find k<=n orthogonal vectors (principal components) that can be best used to represent the data.

Steps:

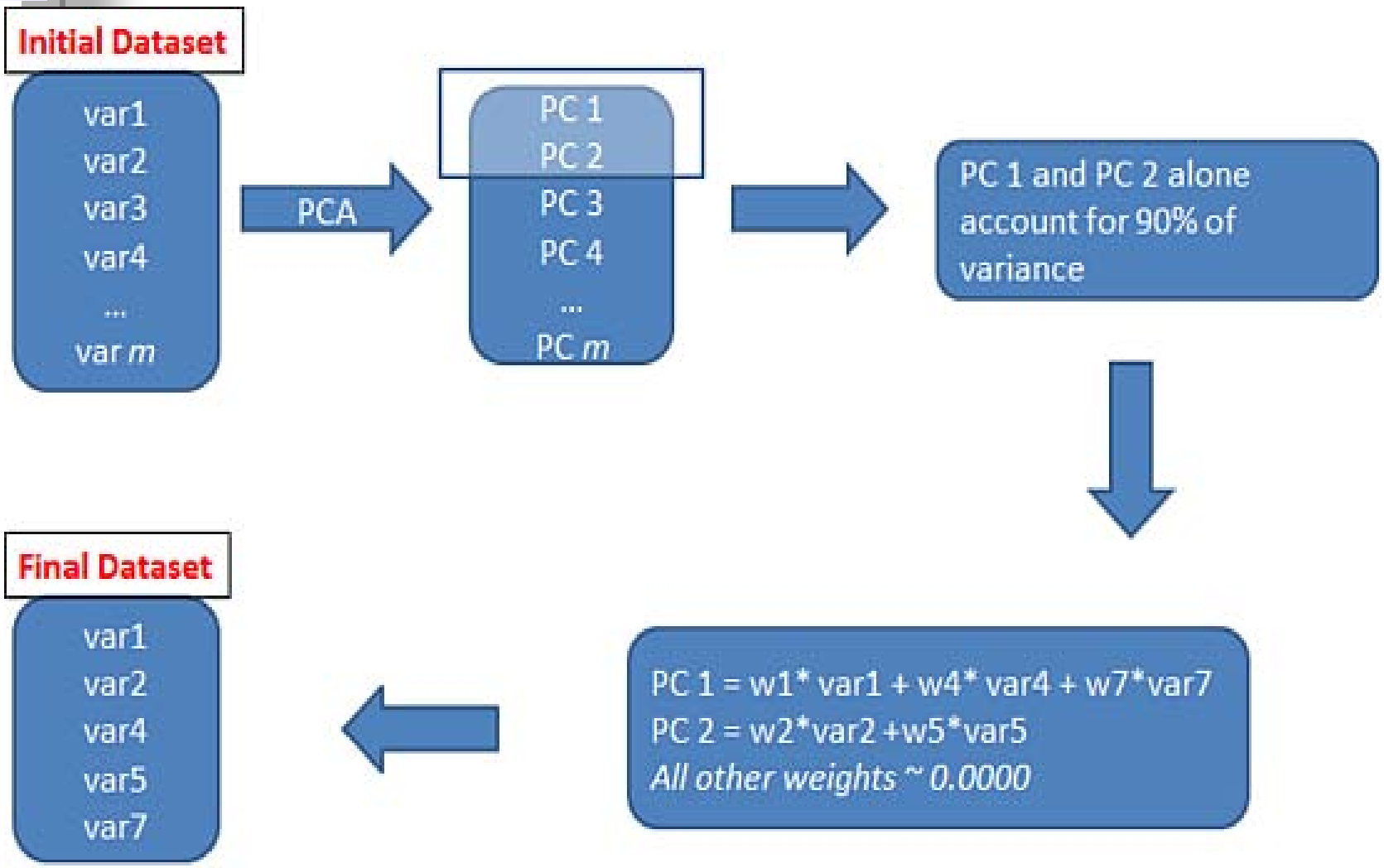
- Normalize input data:
- Compute k orthonomal vectors i.e., principal components
- Principal componets are sorted in decreasing order of significance or strength.
- Reduce data size by eliminating the weak components (low variance).
- Possible to reconstruct a good approximation of original data

Pros:

- Works for numeric data as well as sparse data
- Used for large dimension.

Cons:

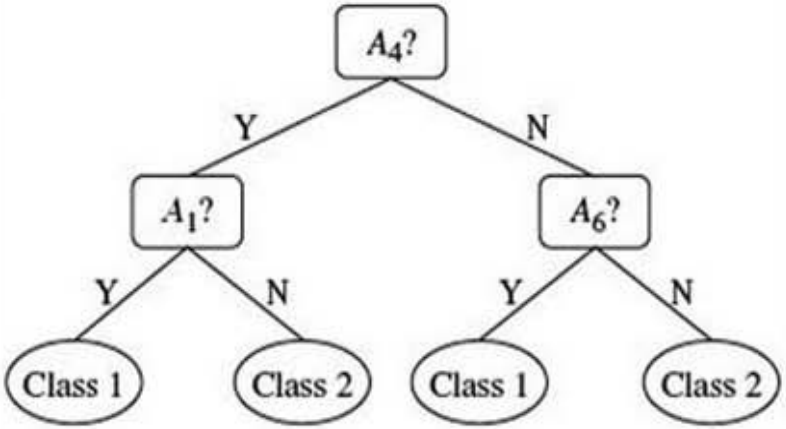
- If the data does not follow a multidimensional normal distribution, pca may not give best principal components.


Initial Dataset
var1
var2
var3
var4
...
var m
PCA
PC 1
PC 2
PC 3
PC 4
...
PC m
PC 1 and PC 2 alone account for 90% of variance
Final Dataset
var1
var2
var4
var5
var7
PC 1 = w1* var1 + w4* var4 + w7*var7
PC 2 = w2*var2 +w5*var5
All other weights ~ 0.0000

Attribute Subset Selection

- Practice of selecting a subset of most consequential attributes for utilizing in model construction.
- Data encloses many redundant or extraneous attributes.

1. Stepwise forward selection:
   Begins with an empty set of attributes.
   Best among the original attributes is determined and added to the reduced set.
   Iterate and add the best from remaining original attributes to reduced set.
2. Stepwise backward elimination:
   Begins with full set of attributes
   Removes worst (least significant) attribute with each step.
3. Combination of forward selection and backward elimination:
   Every step selects the best attribute and removes the worst.
4. Decision tree induction
   Employed to construct a flowchart like structure and each non-leaf node is a test on attribute whereas the external node is a prediction. Algorithm selects best attributes.

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Numerocity Reduction

- Data replaced or estimated by alternative, smaller data representation or fit data into models.
- Parametric model: store model parameter instead of actual data
- Non-parametric (clustering sampling, histogram): storing reduced representation of the data
- Original data volume is replaced into smaller forms of data representation

Normalization

- Transforming all the variables in the data to the same range.
- Heterogenous data with different units needs to be normalized.

Data Scrubbing

- Procedure of modifying or removing incomplete, incorrect, inaccurately formatted, or repeated data in a *database*.
- Objective : make the data more accurate and consistent.
- Performed as batch processing through scripting or data wrangling tools.

Handling missing values

- Important part of the data munging process.
- Incomplete observations adversely affect the operation of machine learning algorithms.

➢ Data imputation is one such procedure – it is the process of filling in missing values based on other data.

➢ Common imputation methods of dealing with unknown or missing values include:

- Removing observations containing one or more unknown values
- Filling in unknown values

1. with most frequent values
2. by exploring correlations
3. by exploring similarities between cases

Feature Extraction

- Storing the extracted features for all subjects in a single project-specific data folder.
- Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant.
- It is a dimensionality reduction process .
- Feature Extraction scripts can be developed using the Sun-Grid engine for speeding up analyses.

Denoising

- Denoising is a signal processing method that extract signal from a mixture of signal and noise thus preserving the useful information.
- Denoising is more significant than any other tasks in voice, video and image processing, analysis and applications.
- Preserving the details of an image, voice or video and removing the random noise as far as possible is the goal of denoising approaches.

Denoising Techniques

Spatial Filtering:

- A traditional way to remove noise from image data.
- Commonly used to clean up the output of lasers, removing aberrations in the beam due to imperfect, dirty or damaged optics.

Spatial filters can be further classified into:

➢ **Linear Filters**:

- Process time-varying input signals to produce output signals, subject to constraint of linearity.
- Linear filters tend to blur sharp edges, destroy lines and other fine image details, and perform poorly in the presence of signal-dependent noise.

➢ **Mean Filter**:

- Acts on an image by smoothing it; reduces intensity variation between adjacent pixels.
- Image corrupted with salt and pepper noise is subjected to mean filtering
- Useful when only a part of the image needs to be processed.

Denoising Techniques contd.

Wavelet Transforms

- Mathematical functions that analyze data according to scale or resolution.
- Studying a signal in different windows or at different resolutions.
- In a large window, gross features in signal can be noticed, but if viewed in a small window, only small features can be noticed.
- Wavelets provide some advantages over Fourier transforms.

LMS Adaptive Filter:

- Capable of denoising non-stationary images
- automatically tracks an unknown circumstance or when a signal is variable with little apriori knowledge
- Does a better job of denoising images compared to the averaging filter
- Works well for images corrupted with salt and pepper type noise.

Sampling:

Data sampling provides a high-quality representative data set for use, ensuring that the three cornerstones for optimal user experience in data wrangling design are delivered to the business users.

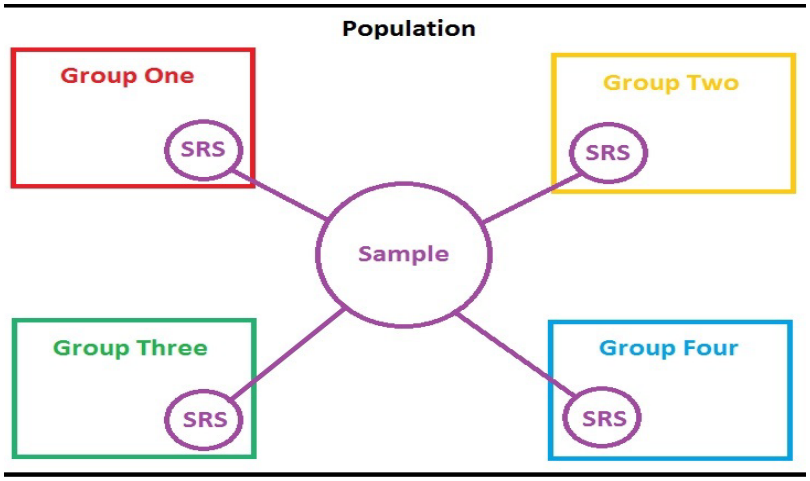
- Facilitates high-performance response times and feedback when working with data sets.
- Ensures engaging user experience that keeps pace with speed of thought for uninterrupted concentration.
- Secures a positive user experience regardless of data volumes
- Focusses on the data's complexity rather than size

Stratified Sampling

- Process used in market research Involves dividing the population of interest into smaller groups, called strata.
- Samples are then pulled from these strata, and analysis is performed to make inferences about the greater population of interest.

When is Stratified Sampling used?

- Target population of interest is significantly heterogeneous
- Highlight specific subgroups within population of interest.
- Observe the relationship(s) between two or more subgroups.
- Create representative samples from smallest, most inaccessible subgroups of the population



Steps to Perform Stratified Sampling

Step 1: Divide the population into smaller subgroups, or strata, based on the members' shared attributes and characteristics.

Step 2: Take a random sample from each stratum in a number that is proportional to the size of the stratum.

Step 3: Pool the subsets of the strata together to form a random sample.

Step 4: Conduct analysis.

THANK YOU !!!