



The
Center of
Applied
Data Science

Big Data and Apache Hadoop

November 2019



Contents

1. 5Vs of Big Data
2. Types of Data
3. Introduction to Apache Hadoop
4. Principles of Hadoop
5. Hadoop Ecosystem
6. Use Cases of Apache Spark

5Vs of Big Data

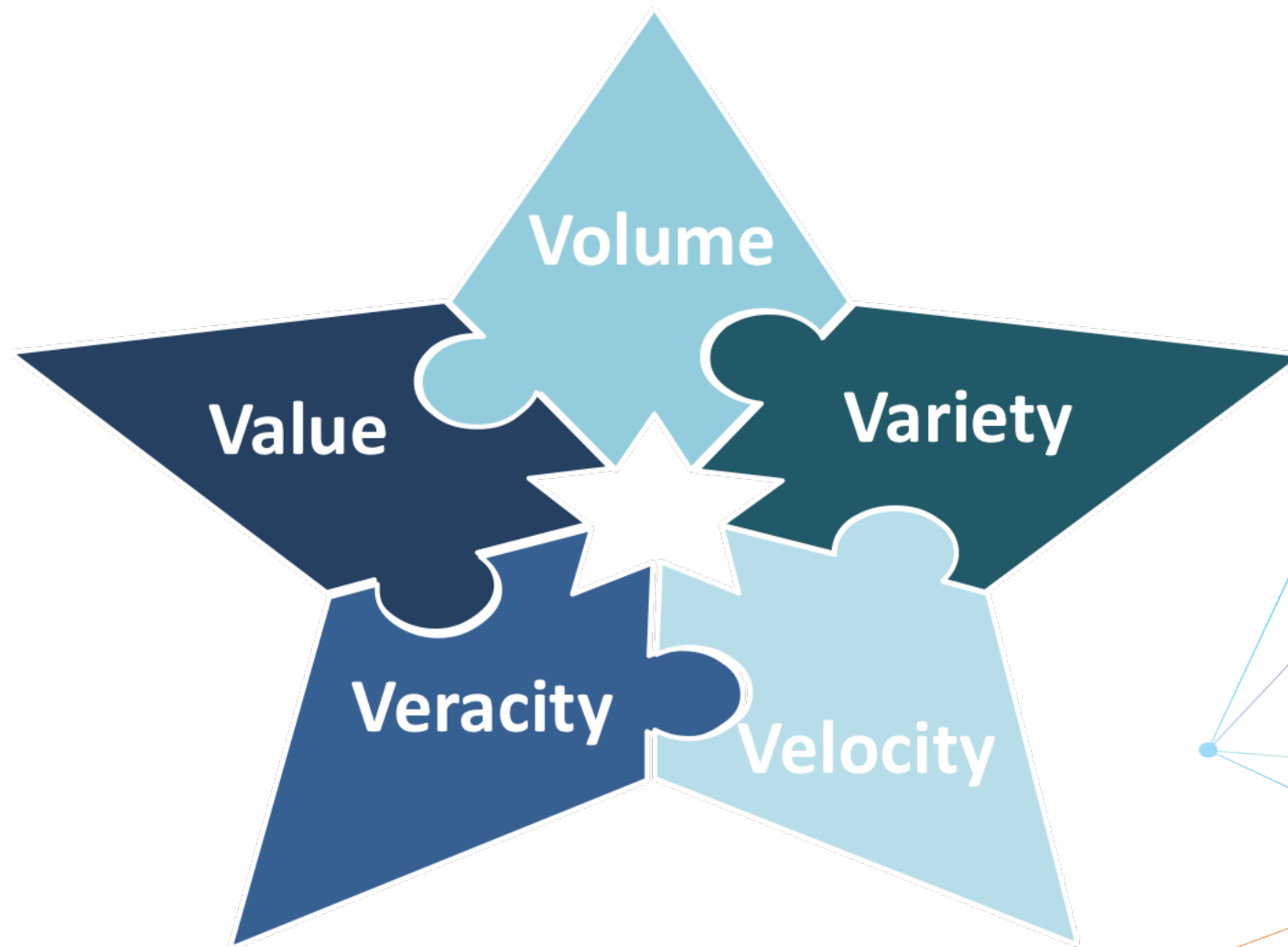




The
Center of
**Applied
Data Science**

5Vs of Big Data

Definition





5Vs of Big Data

Definition

Volume

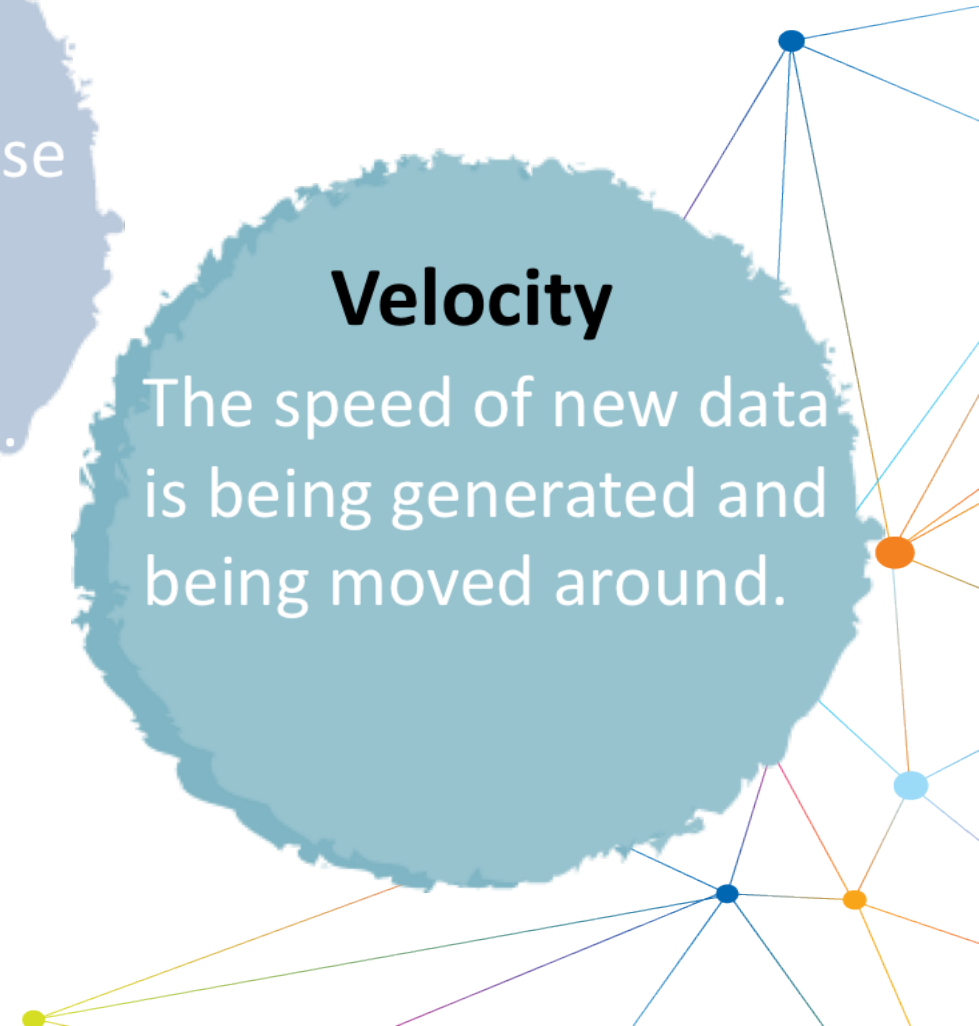
Vast amounts of data (Zettabytes/ Brontobytes) generated every second.

Variety

Different types of data we can now use (structured, semi-structured and unstructured data).

Velocity

The speed of new data is being generated and being moved around.





5Vs of Big Data

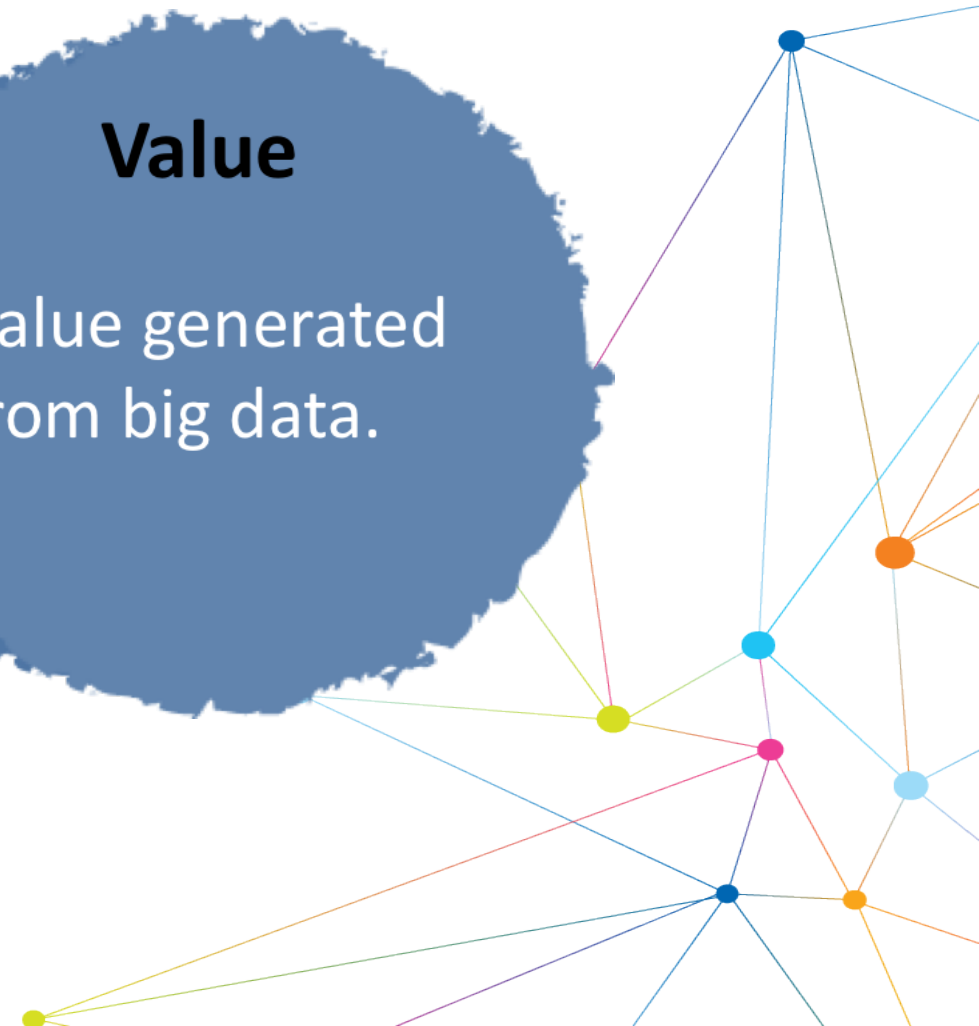
Definition

Veracity

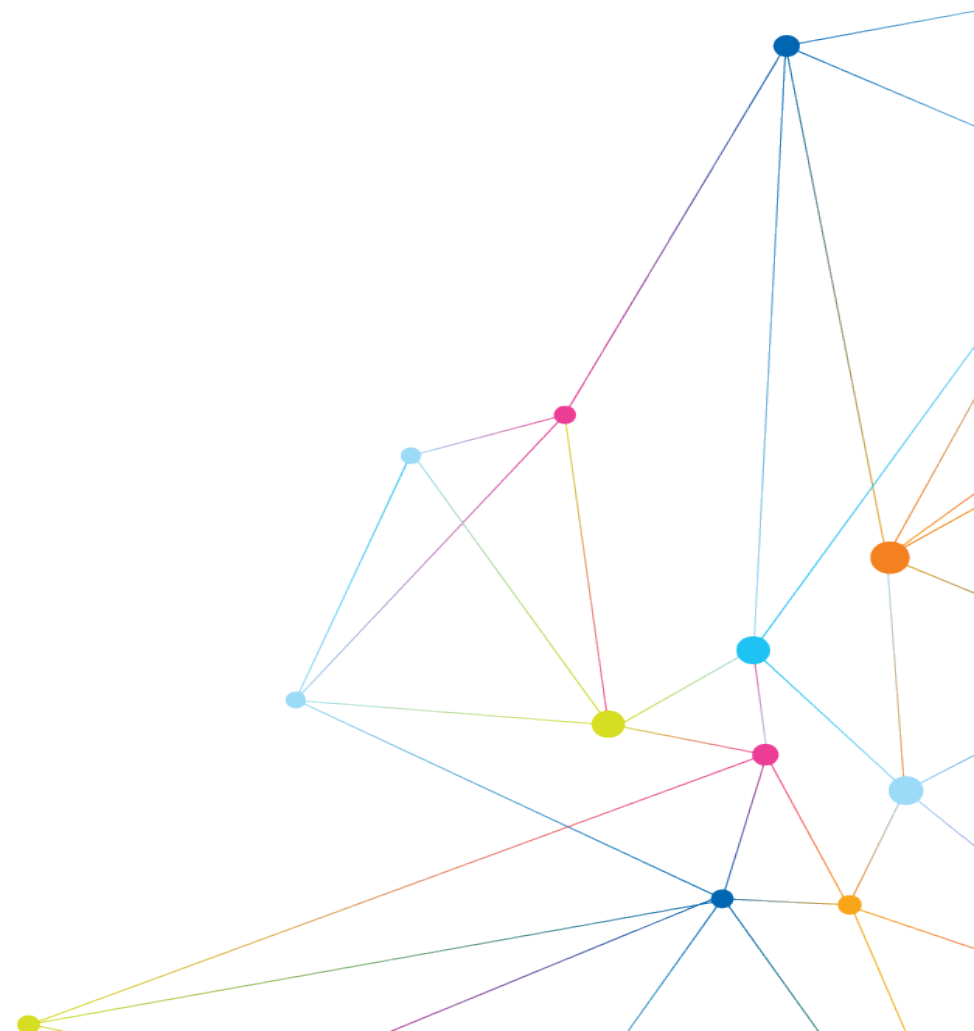
Able to use data
regardless of the
quality and accuracy.

Value

Value generated
from big data.



Types of Data





Structured

- Stored in **databases**.
- Organized in rows and columns.
- **Example:** Data received from web logs and sensors.

Semi-Structured

- Data that is not stored in traditional database, but being stored in **certain organizational way**.
- **Example:** NoSQL documents

Unstructured

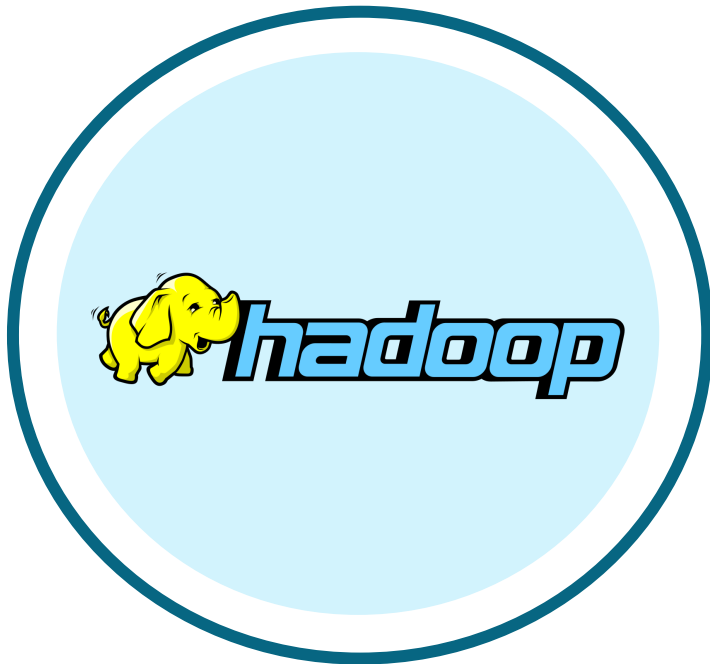
- Data that **do not have clear format** in storage.
- **Example:** Pictures uploaded online, YouTube videos, Text messages sent to social media



The
Center of
**Applied
Data Science**

Introduction to Apache Hadoop



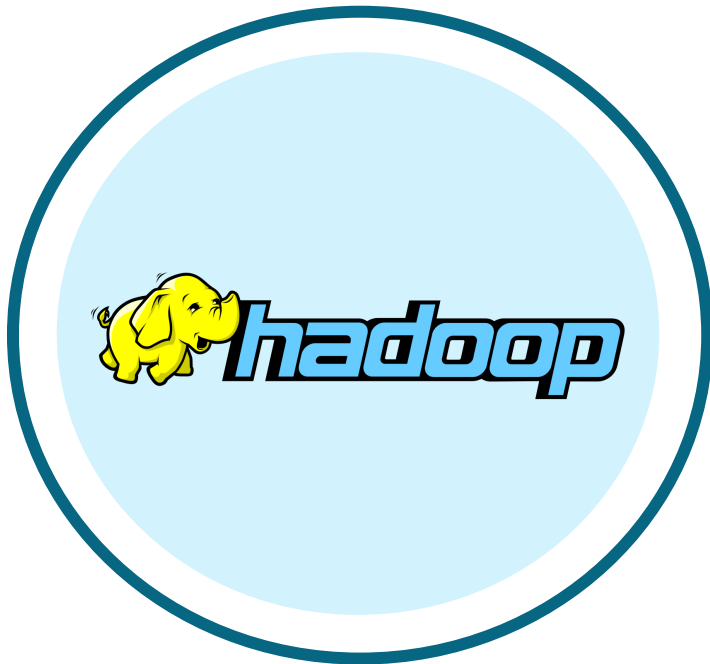


Overview

Open source programs and frameworks which can be used as the backbone of the big data operations.

Advantages

- Scalability
- Reliability
- Flexibility



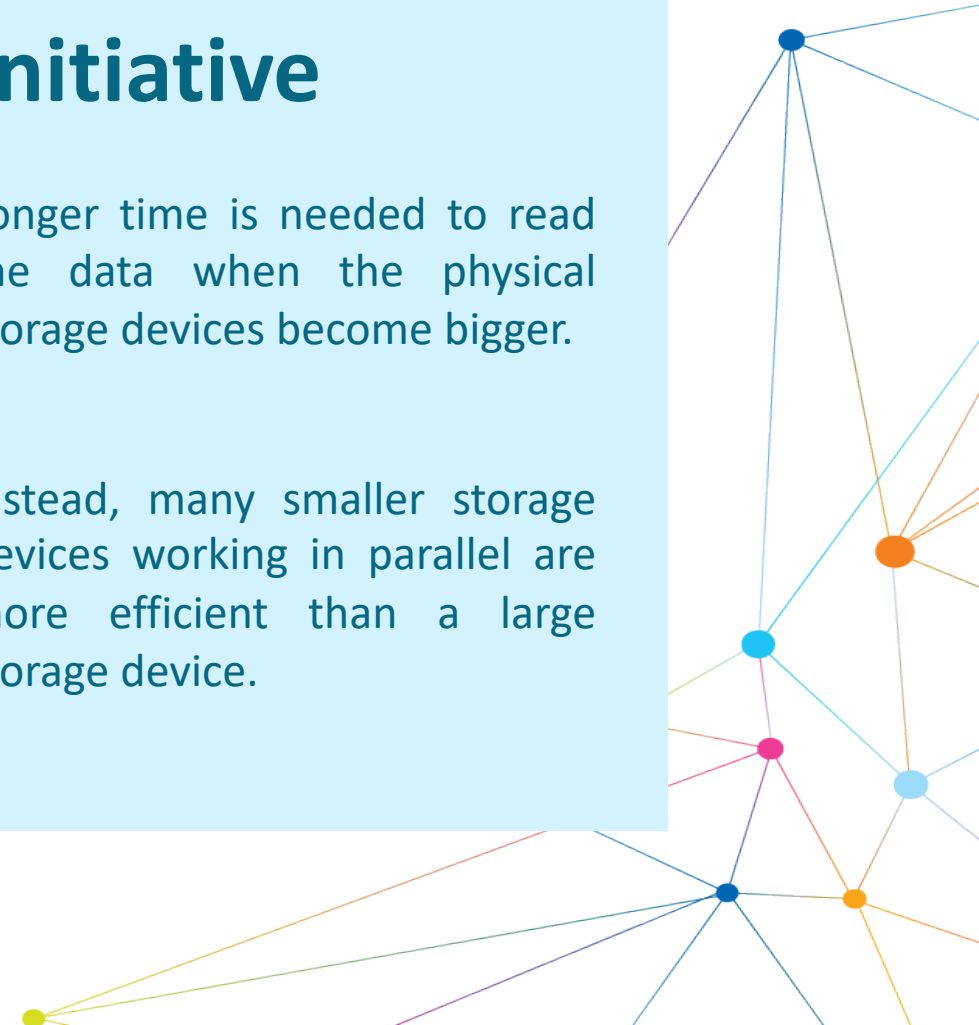
Initiative



Longer time is needed to read the data when the physical storage devices become bigger.



Instead, many smaller storage devices working in parallel are more efficient than a large storage device.





The
Center of
**Applied
Data Science**

Principles of Hadoop

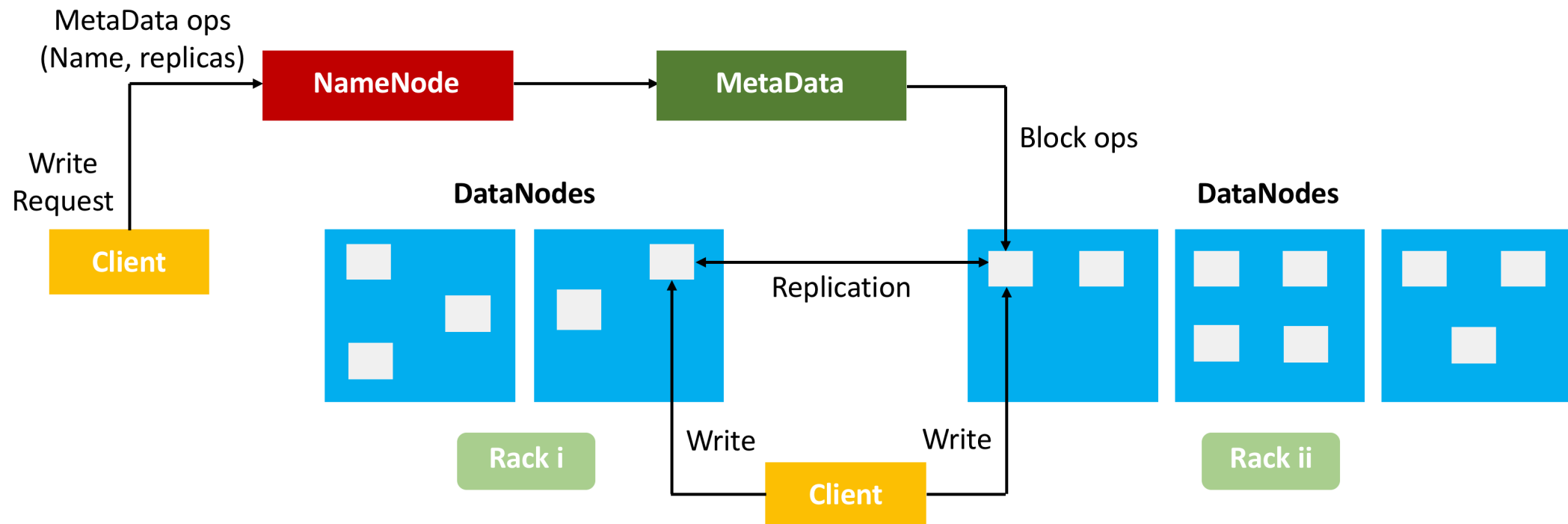


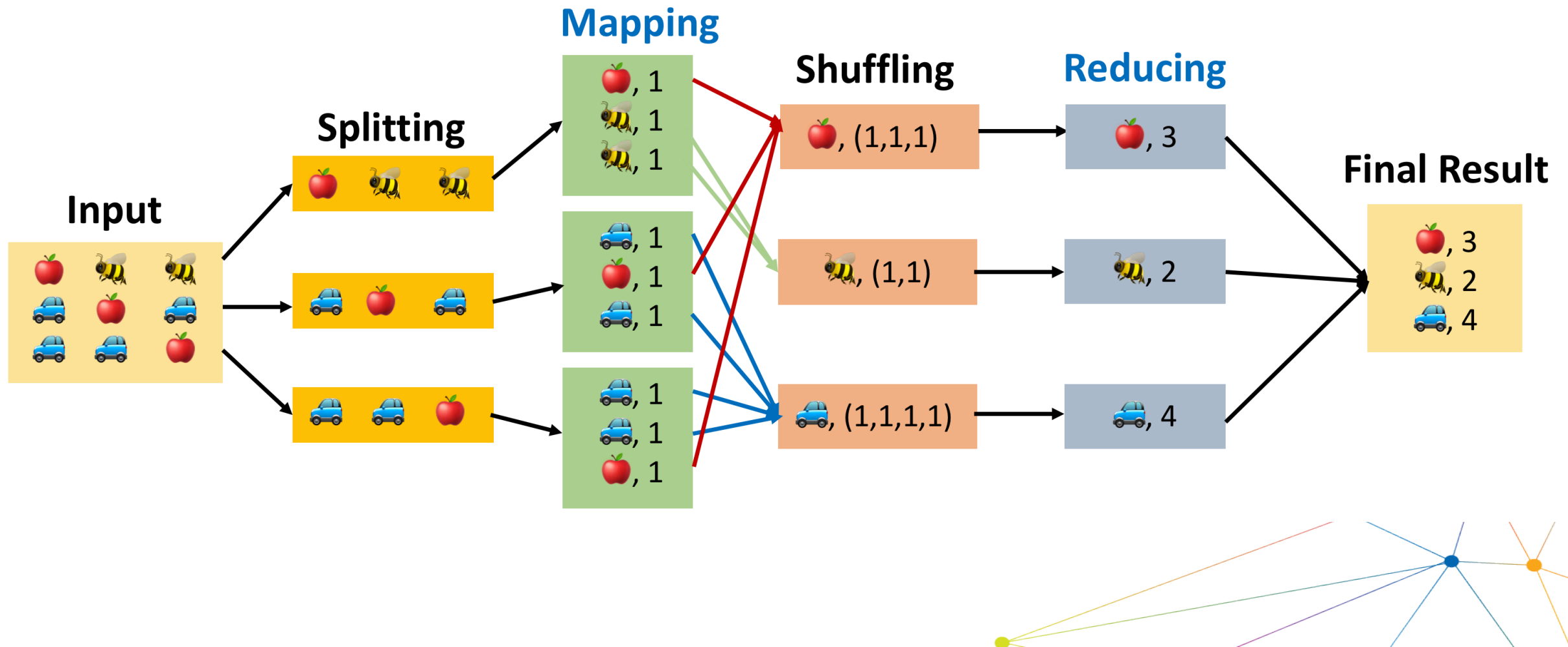


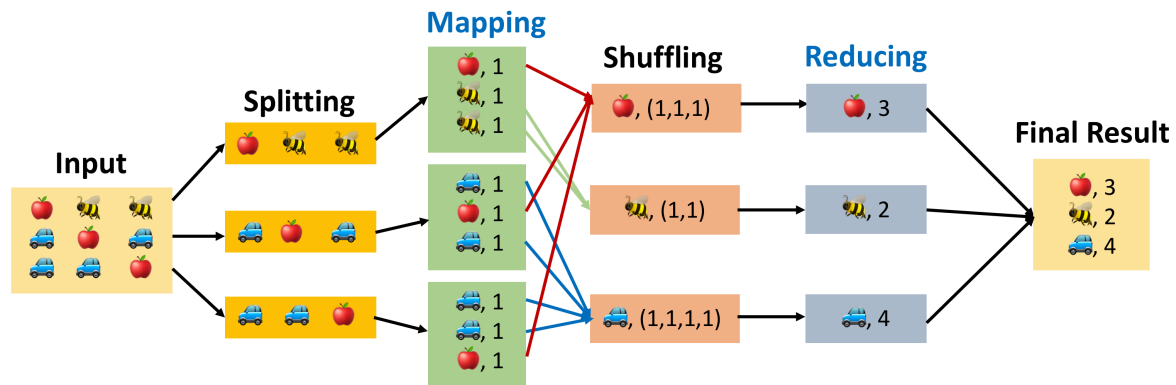
Overview

- **Reliable** architecture to store very large files in Hadoop cluster.
- **Store less number of large files** rather than huge number of small files.
- **Fault tolerance.**
- **High throughput** by providing data access in parallel.

HDFS Architecture





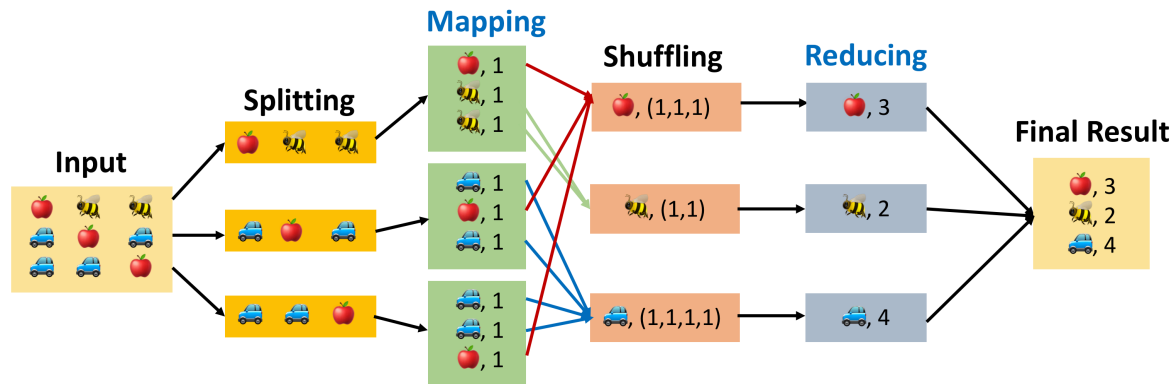


Terminologies

Job	Complete process from input to final output
Task	A part of the job executed on a slice of data
JobTracker	Master node to manage the jobs and resources
TaskTracker	Agent deployed in each machine to run MapReduce

Task of Mapper

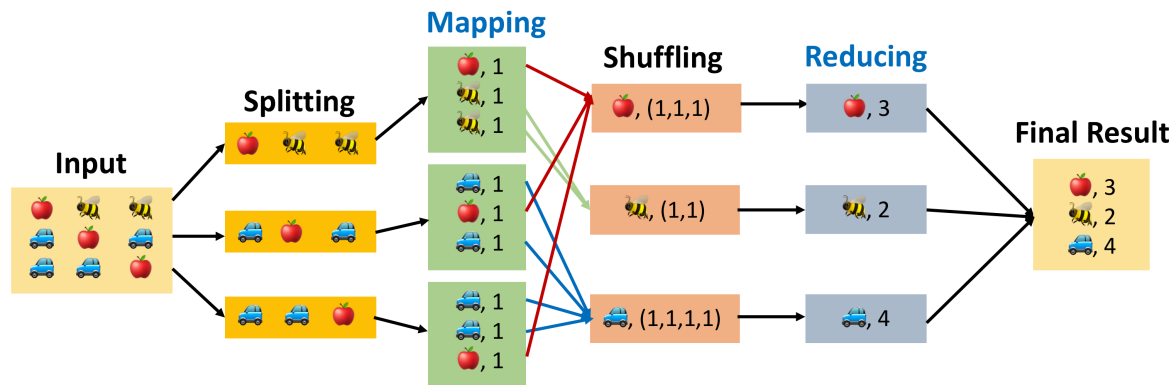
- The input is mapped into Key-Value (KV) pair.
- For example $\langle \text{🍏}, 1 \rangle$ is in the format of $\langle \text{key}, \text{value} \rangle$.



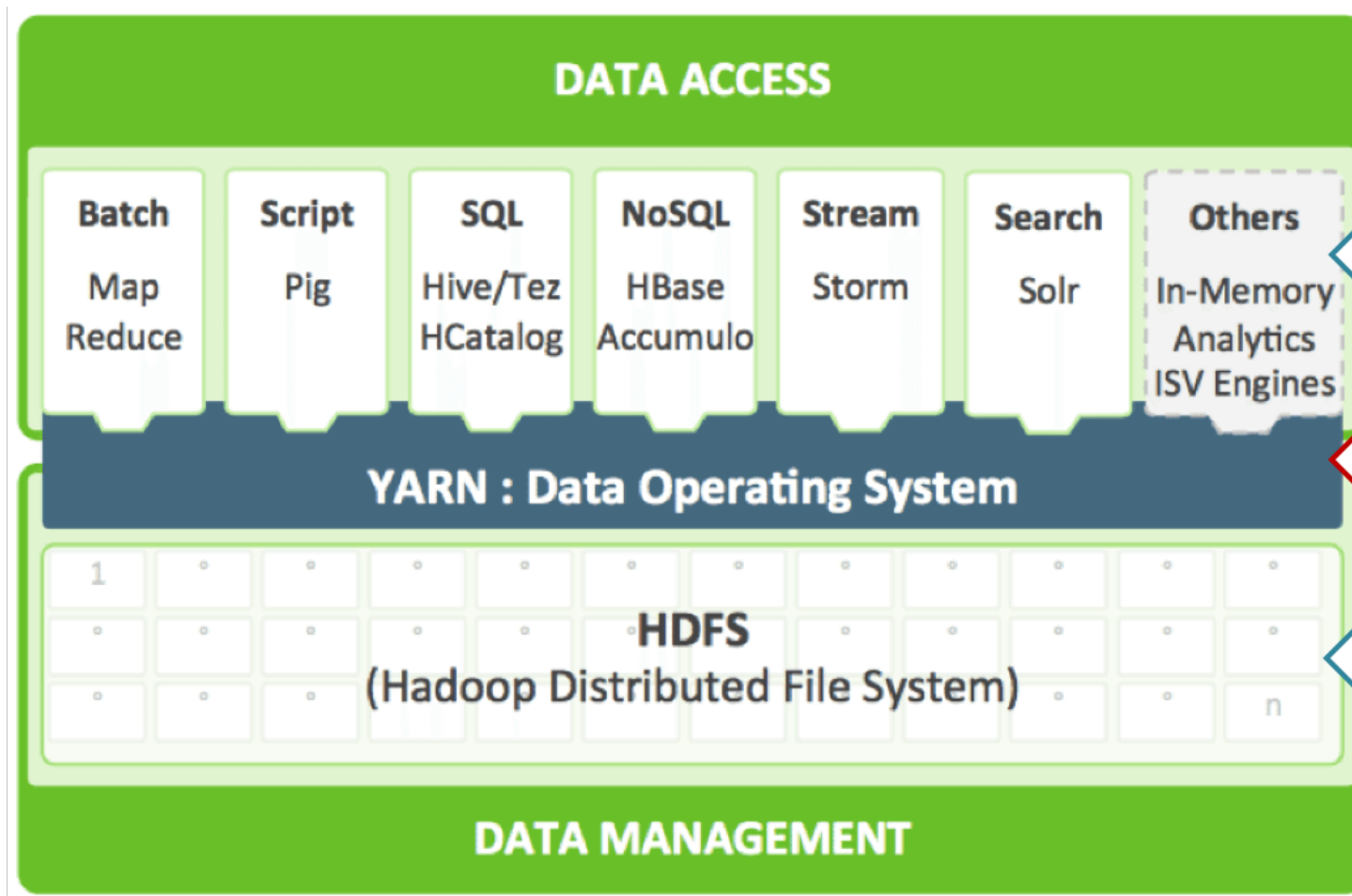
Intermediate Process

- Mapper output undergoes shuffle and sorting.
- The intermediate data would be stored in local file system without having replications in other nodes.

Task of Reducer



- Started only after all the mappers have completed their operations.
- Perform mathematical operations (such as aggregation/summation).
- User could define function to meet custom business logic.
- The output of Reducer is stored in HDFS.



More platforms than just MapReduce

YARN for better resource utilization

Common HDFS Foundation



The
Center of
**Applied
Data Science**

Hadoop Ecosystem





The
Center of
Applied
Data Science

Hadoop Ecosystem

Apache Spark



Overview

- Open source cluster computing framework that is suitable for large-scale data processing.
- Provides machine learning projects, batch processing, near real-time processing, and graph analysis.



The
Center of
Applied
Data Science

Hadoop Ecosystem

Apache Spark



Application

- Suitable for large-scale Data Science use cases.
- Able to run on Hadoop, Amazon AWS cloud, and different databases such as Cassandra, Amazon Dynamo DB etc.



Overview

- Virtual data warehouse software to perform MapReduce based SQL engine that runs on top of Hadoop.
- Apache Hive employs HiveQL (SQL-like query language) to access the files stored in Apache HDFS or other data storage system such as Apache Hbase.



Application

- Suitable to build data warehouse without requiring programmers to write complex MapReduce code.
- A real world application is the friend recommendation system on Facebook. Recommendation system has two characteristics:
 - Require high volume of input data.
 - Outputs/Recommendations do not change frequently.



Overview

- Distributed, scalable, and multi-level big data store on top of Hadoop and HDFS.
- NoSQL database used for real-time data streaming due to the two advantages:
 - Able to provide fast and random read-writes.
 - Able to work well with sparse data due to the column-oriented property



Application

- Suitable for random, real-time read/write access to big data.
- Real world applications of Apache Hbase:
 - Helps Facebook to perform real-time analytics, such as counting Facebook likes and for messaging.
 - Helps Financial Industry Regulatory Authority (FINRA) and Pinterest to store graphs.
 - Helps Flipboard to personalize the content feed for the users.



Overview

- Platform for analyzing large data sets.
- Apache Pig employs Pig Latin for queries and data manipulation.
- Apache Pig has competitive advantages to perform more complex data manipulation queries by providing:
 - Nested data types like Maps, Tuples, and Bags.
 - Support to major data operations like Ordering, Filters, and Joins.



Application

- Suitable for constructing scheduled job. Hence, it is appropriate for automated batch jobs that move data between HDFS and other systems.
- Suitable to read data from the databases reside in Hadoop that are not structured with Apache Hive metadata schemas.



Overview

- A tool for automating the transfer process of bulk data between Apache Hadoop and structured datastores efficiently.
- Able to execute the data transfer in parallel.



Overview

- A tool to collect, aggregate, and transport large amounts of streaming data from variety of sources in both real-time and batch mode to a centralized data store (e.g. HDFS).



The
Center of
Applied
Data Science

Hadoop Ecosystem

Apache Flume



Application

- Suitable to import huge volumes of event data generated by websites such as Facebook, Twitter, Amazon, and Flipkart in real-time.



Overview

- A tool for data streaming and processing applications and it is exceling for stateful streaming applications at any scale.
- Provides real-time processing, machine learning projects, batch processing, and graph analysis.



Application

- Apache Flink is able to run on third-party data sources such as Amazon Kinesis Streams, Elasticsearch, Cassandra, and Twitter Streaming API.
- The introduction of ACID into data Artisans platform reinforces position of Apache Flink as the integration hub for the real-time large financial and eCommerce organizations.



Overview

- A low latency high performance SQL like queries engine to query data that stored on Hadoop clusters (e.g. HDFS, Apache Hbase) in real-time.
- Impala shares the same SQL syntax (Hive SQL), ODBC driver, metadata, and user interface (Hue Beeswax) as Apache Hive. Hive users can then use Impala with little setup overhead.



The
Center of
Applied
Data Science

Hadoop Ecosystem

Apache Impala



Application

- Suitable for the interactive applications that require complicated queries to react relatively fast. It allows users to obtain the outputs to the unexpected questions (complicated queries) in seconds or at most a few minutes.



Overview

- Designed to build a central data backbone for a large organization with a single cluster for processing ingests data in real-time.
- A single Kafka broker can handle hundreds of megabytes of reads and writes per second from thousands of clients.



Application

- Suitable for publish-subscribe messaging. Users can publish and subscribe to information as and when they occur.
- Suitable to manage the variety of use cases commonly required for a Data Lake.
- Able to render streaming data through a combination of Apache Hbase, Apache Storm, and Apache Spark systems.



Overview

- A real-time computational system for accepting high volume data coming in high velocity, possibly from various sources.
- Easy to implement and can be integrated with any programming language.



Application

- Suitable for applications that primarily focused on stream processing and CEP-style processing.
- Apache Storm has the advantage of broader language support over Apache Spark.



Overview

- A schema-free SQL query engine for Hadoop, NoSQL, and cloud storage.
- Does not depend on Hadoop as Drill does not use MapReduce job internally. Drill has its own distributed processing service called DrillBit.



Application

- Can be used to join data from multiple datastores with just a single query.
- Can be used to connect between standard BI/analytics tools and non-relational datastores by leveraging Apache Drill's JDBC and ODBC drivers.



Apache Arrow

Overview

- Built by the lead developers of many Apache projects.
- A component used to exchange data with low overhead and hence accelerating the data analytics.
- Apache Arrow is extremely important for Python and R communities as it provides data interoperability between the two communities with big data systems (which largely run on the JVM).



Apache Arrow

Application

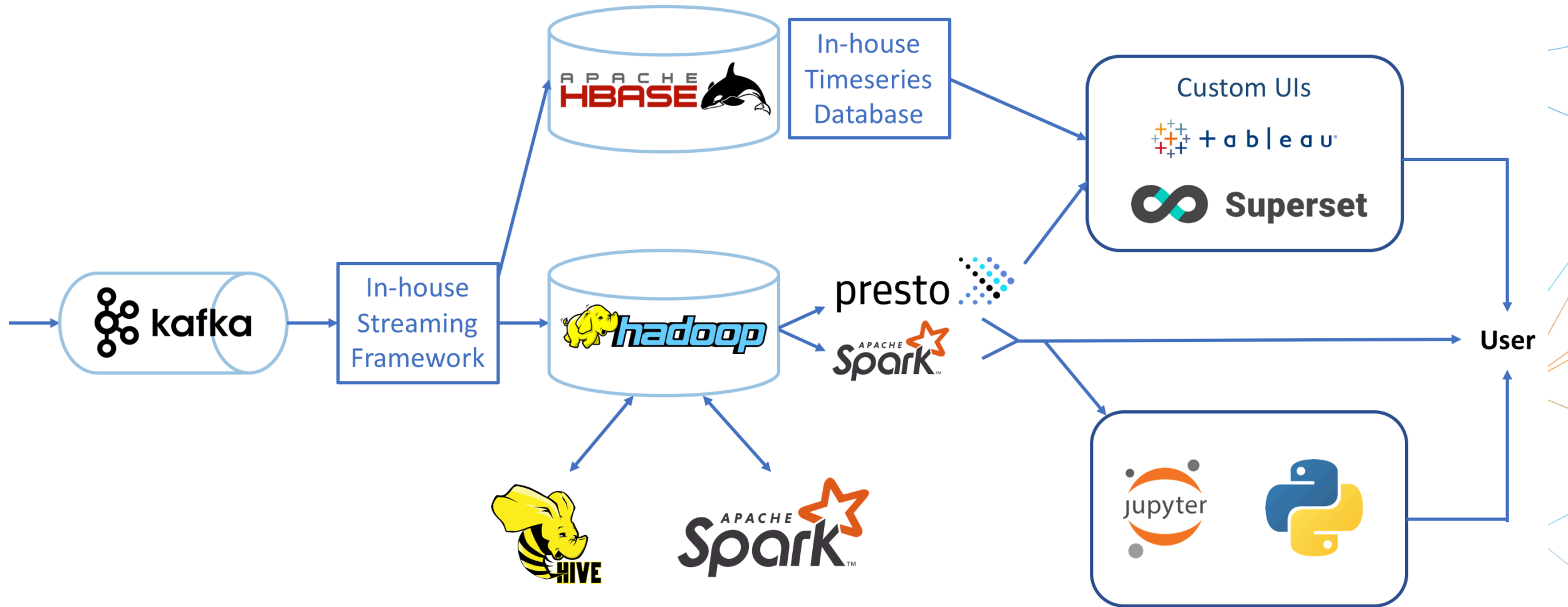
- Used to reduce the time spent gathering and processing data. For example:
 - PySpark: IBM measured a 53x speedup in data processing by Python and Spark with the support of Apache Arrow in PySpark.



The
Center of
Applied
Data Science

Hadoop Ecosystem

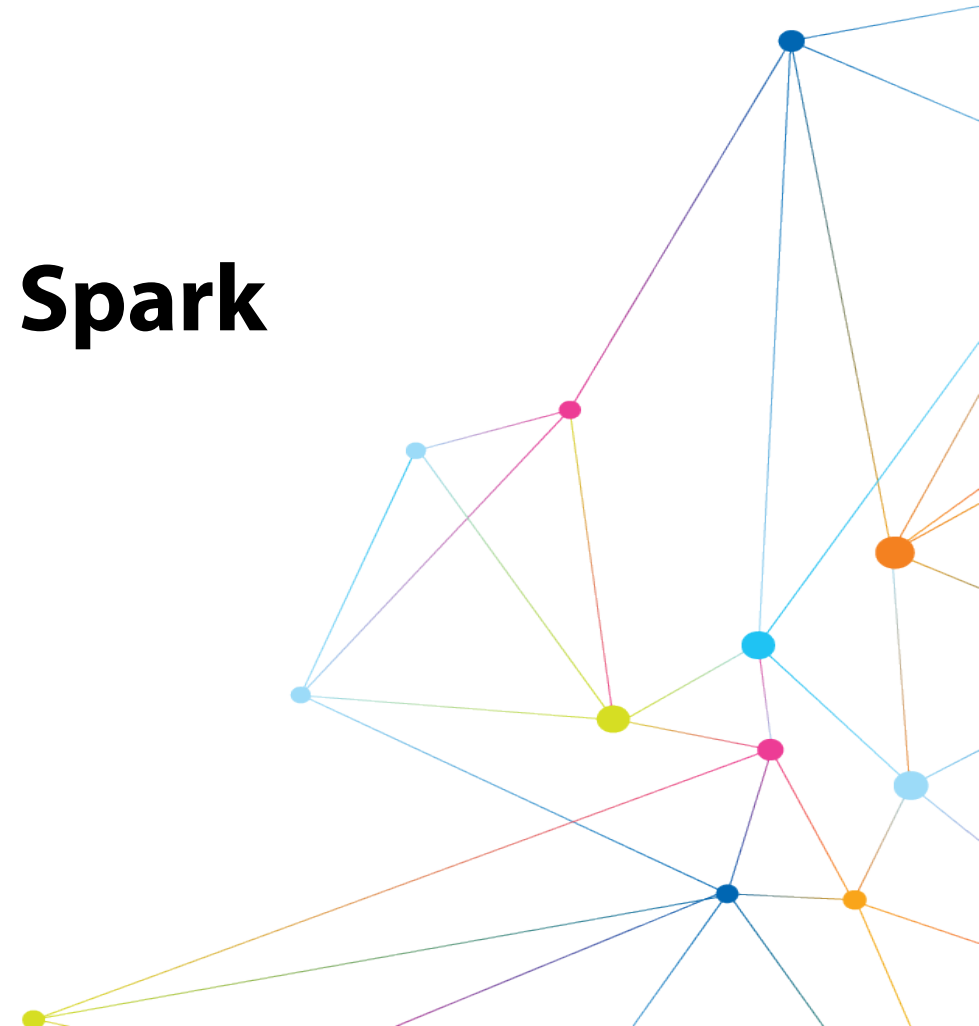
Big Data Technology Stack from Tesla





The
Center of
**Applied
Data Science**

Use Cases of Apache Spark



YAHOO!

INDUSTRY:

Web Services Provider

USE CASE:

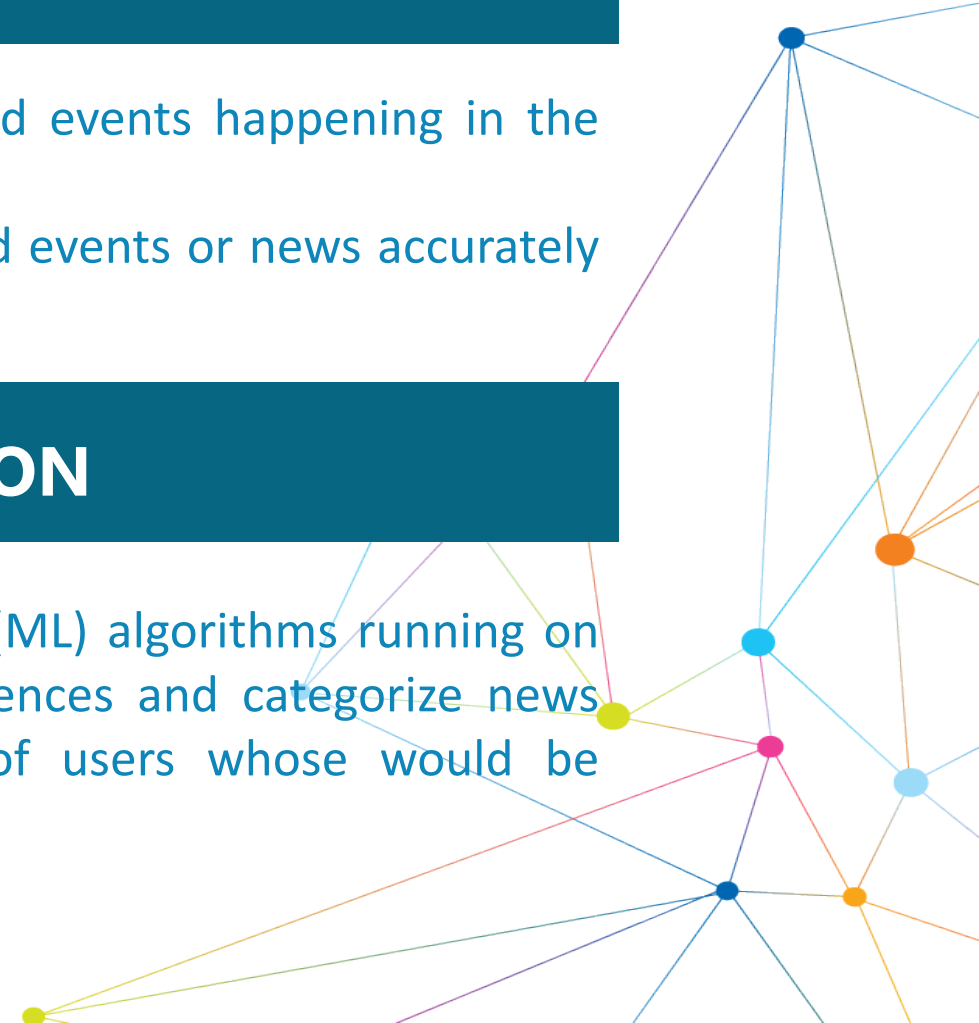
News Pages Personalization

CHALLENGES

- Analyzes users' preferences and events happening in the outside world.
- Relates users to their interested events or news accurately and promptly.

SOLUTION

- Yahoo uses Machine Learning (ML) algorithms running on Spark to analyze users' preferences and categorize news stories based on the types of users who would be interested in reading them.



YAHOO!

INDUSTRY:

Web Services Provider

USE CASE:

Advertisement Analytics with Existing
BI Tools

CHALLENGES

- To make Spark compatible with existing BI tools to view and query the advertising analytic data stored in Hadoop.

SOLUTION

- Spark Shark is compatible with the standard Hive server API and hence there is no issue to work with tools that plugs into Hive (e.g. Tableau).
- With the compatibility, Yahoo is able to query their advertisement visit data interactively.

conviva®**INDUSTRY:**

Online Video Streaming Provider

USE CASE:

Online video optimization and online video analytics

CHALLENGES

- Users expect to have good video quality without much delays.
- Require highly-sophisticated behind-the-scenes technology to ensure a high quality of service by avoiding dreaded screen buffering.

SOLUTION

- Conviva deploys Spark Streaming to analyze the network traffics in real time. Subsequently, the results are fed directly into the video player (e.g. Flash player) to optimize the speeds.



INDUSTRY:

Data Intelligence Company

USE CASE:

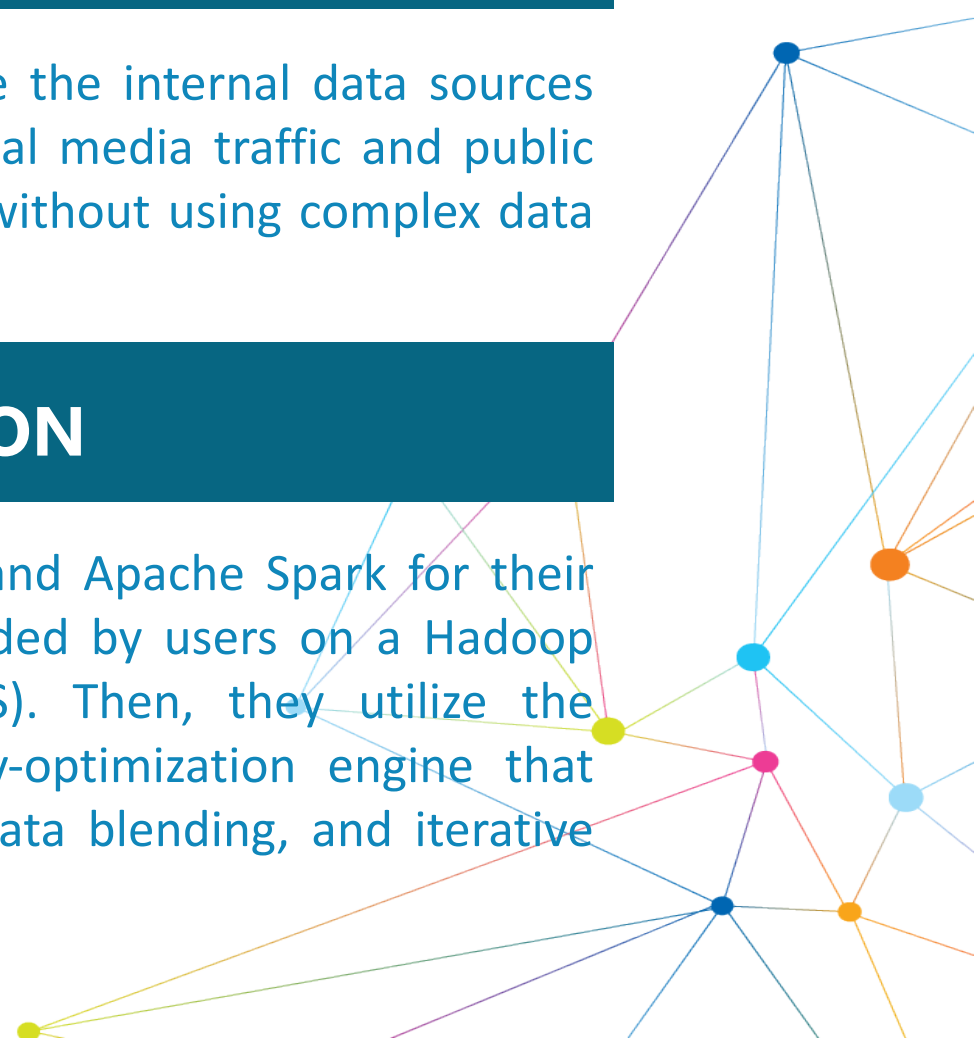
Internal and External Data
Harmonization

CHALLENGES

- Require a platform to integrate the internal data sources with external sources (e.g. social media traffic and public data feeds) for business users without using complex data modeling.

SOLUTION


- ClearStory uses both Hadoop and Apache Spark for their service. They store data uploaded by users on a Hadoop Distributed File System (HDFS). Then, they utilize the Spark's core in-memory query-optimization engine that allows fast data preparation, data blending, and iterative analysis.






Thank you.

 thecads.org

 info@thecads.org

 The Center of Applied Data Science

 thecads.org

 thecadsmalaysia