



APACHE SPARK

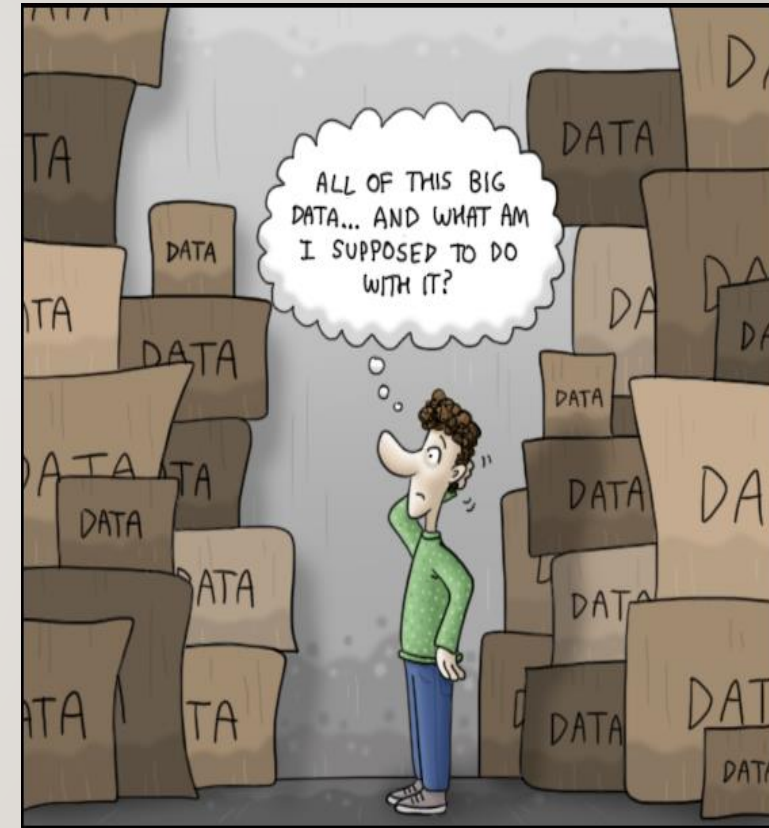
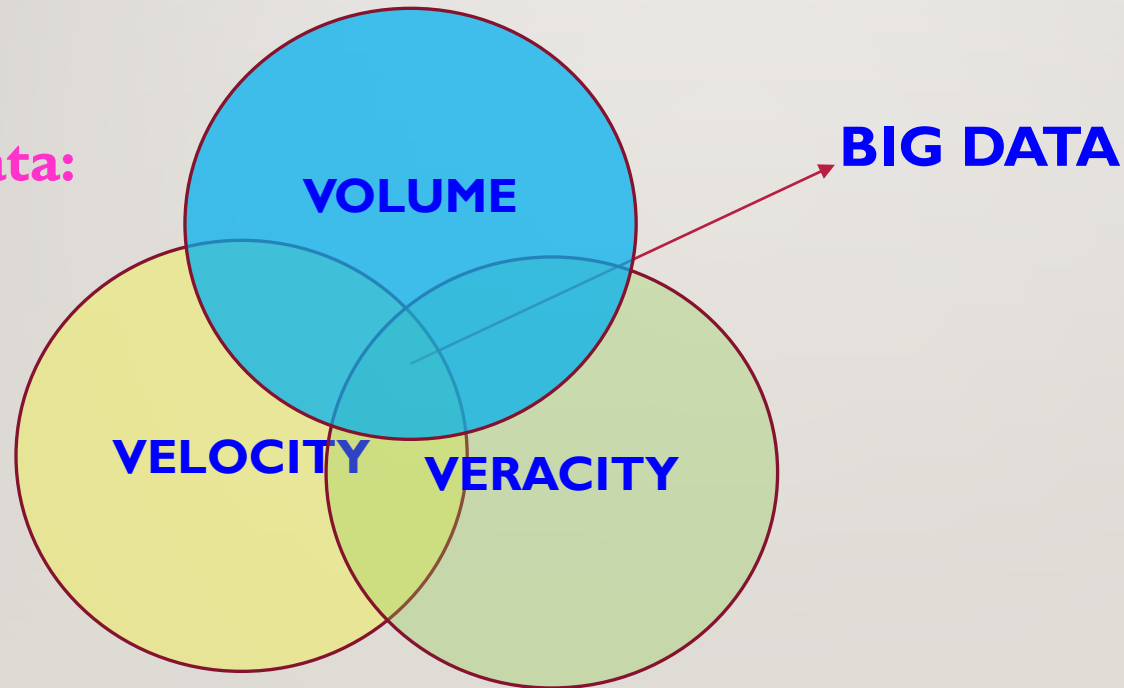
SMRITHI AJIT

BIG DATA WORKSHOP

BIG DATA

“Data that is so big and complex that it is hard to process it”

3Vs of Big Data:



CONCEPT OF HADOOP

OPEN SOURCE

SOFTWARE
PLATFORM

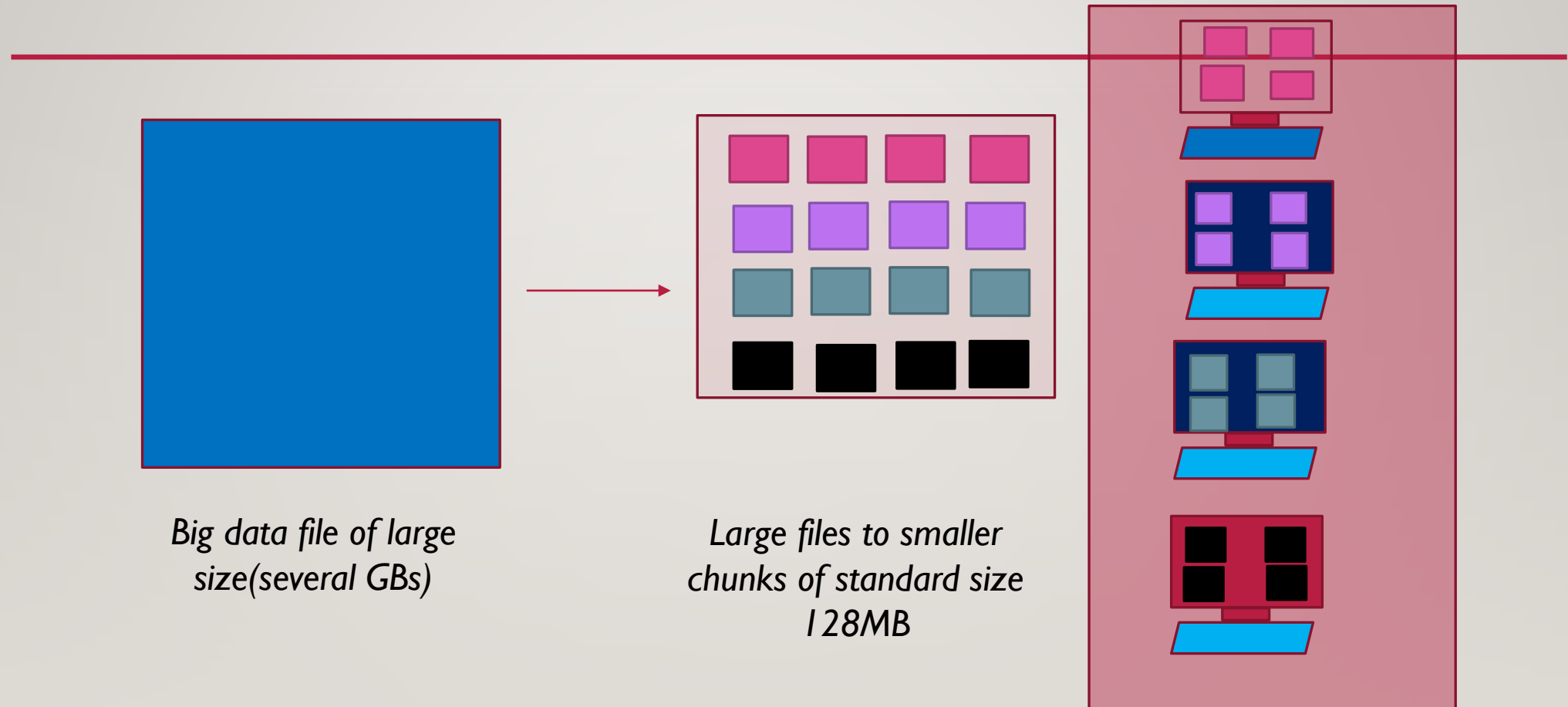
DISTRIBUTED
STORAGE

DISTRIBUTED
PROCESSING

LARGE DATA
SETS

COMPUTER
CLUSTERS

HADOOP DISTRIBUTED FILE SYSTEM(HDFS)



SIMPLE HDFS COMMANDS AND FUNCTION

`hdfs dfs -ls`

List the files on the HDFS

`hdfs dfs -copyToLocal <hdfs source>/filename
<destination>`

Copy files from HDFS to local folder

`hdfs dfs -put <source> <hdfs destination>`

Put files from the local folder to HDFS

`hdfs dfs -mv <source> <destination>`

Move file from one location on HDFS to another

`hdfs dfs -cp <source> <destination>`

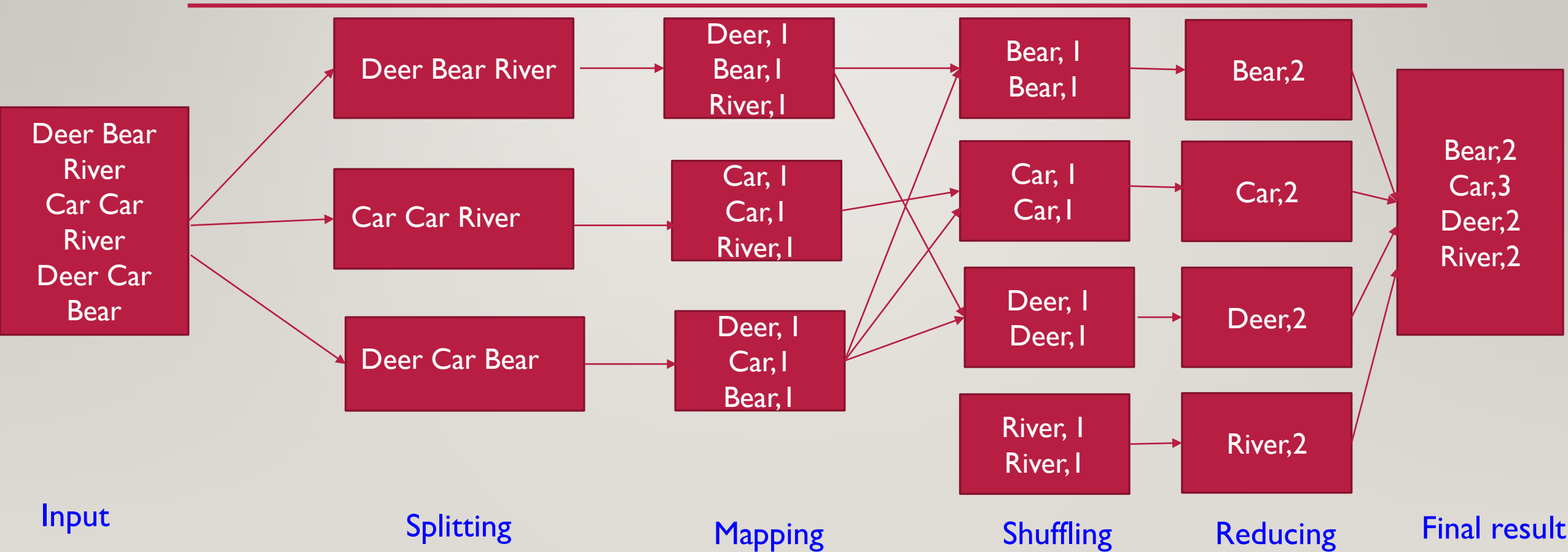
Copy files from one location on HDFS to another



CONCEPT OF DISTRIBUTED COMPUTING AND BIRTH OF SPARK

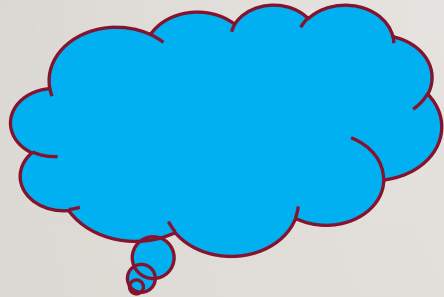
- Google introduced: distributed data, distributed processing and fault tolerance
- MapReduce, a resilient distributed processing framework, which enabled Google to index the exploding volume of content on the web, across large clusters of commodity servers.
- After Google's white paper in 2004, Doug Cutting and Mike Cafarella created Apache Hadoop
- Apache Spark born in 2009, AMPLab at the University of California, Berkeley.
- Spark incubated by Apache Software Foundation in 2013, and promoted early in 2014
- Today it is one of the most active projects managed by the Foundation
- Well-funded corporate backers: Databricks, IBM, and China's Huawei.

CONCEPT OF MAPREDUCE

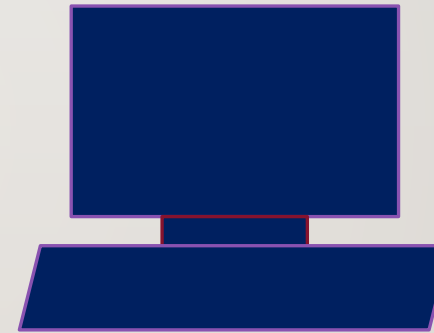


APACHE SPARK

“An open-source big data processing framework built around speed, ease of use, and sophisticated analytics”



Paid and Free services: AWS, Microsoft Azure, Google Colabs, Databricks



On-premise: Local system, Hadoop cluster

HADOOP MAP REDUCE VERSUS APACHE SPARK

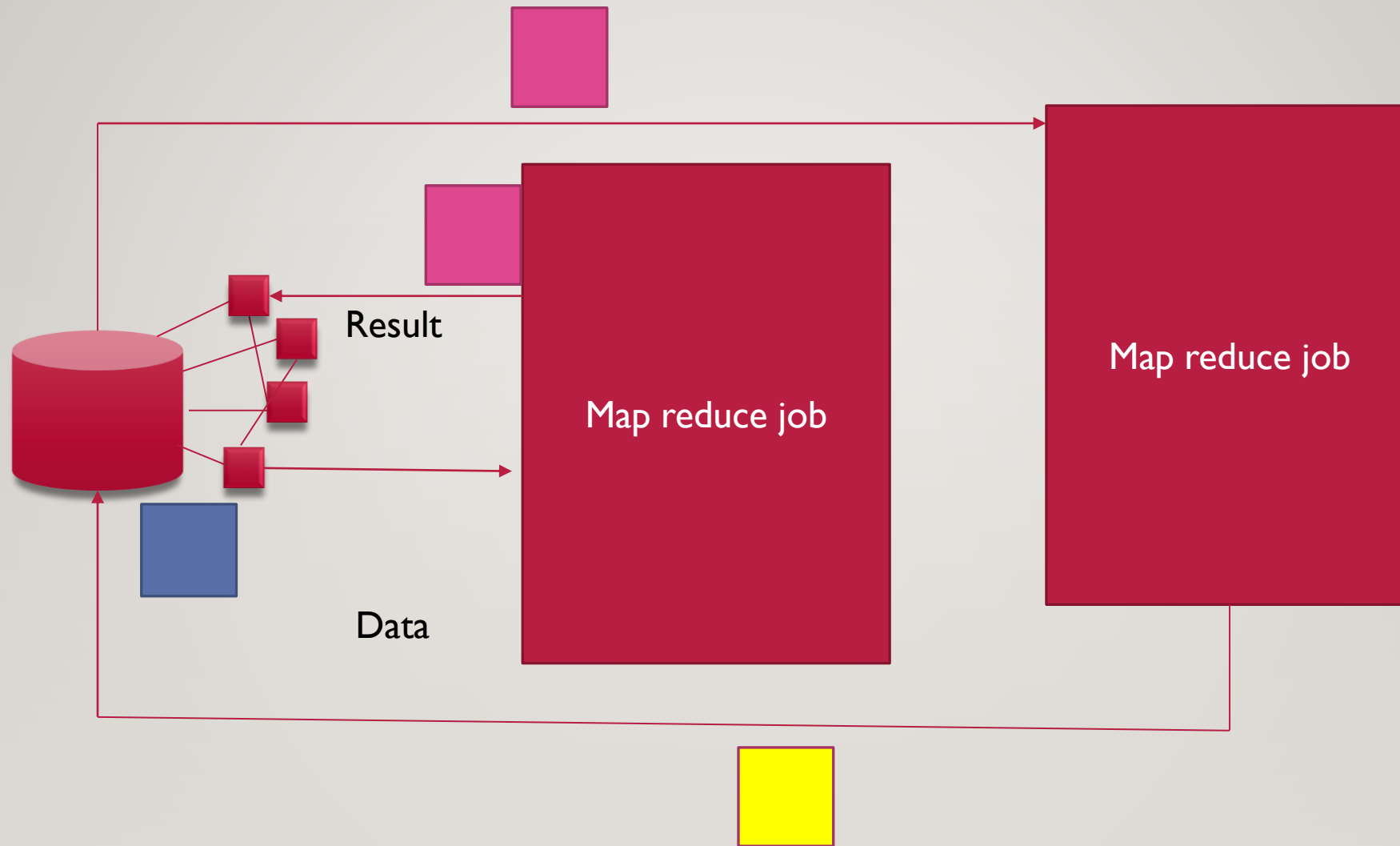
MAP REDUCE

- *Only batch processing*
- *Slower than Apache spark because of I/O latency*
- *Write all intermediate results to disk*
- *Spend 90% of the time on HDFS read write operations*

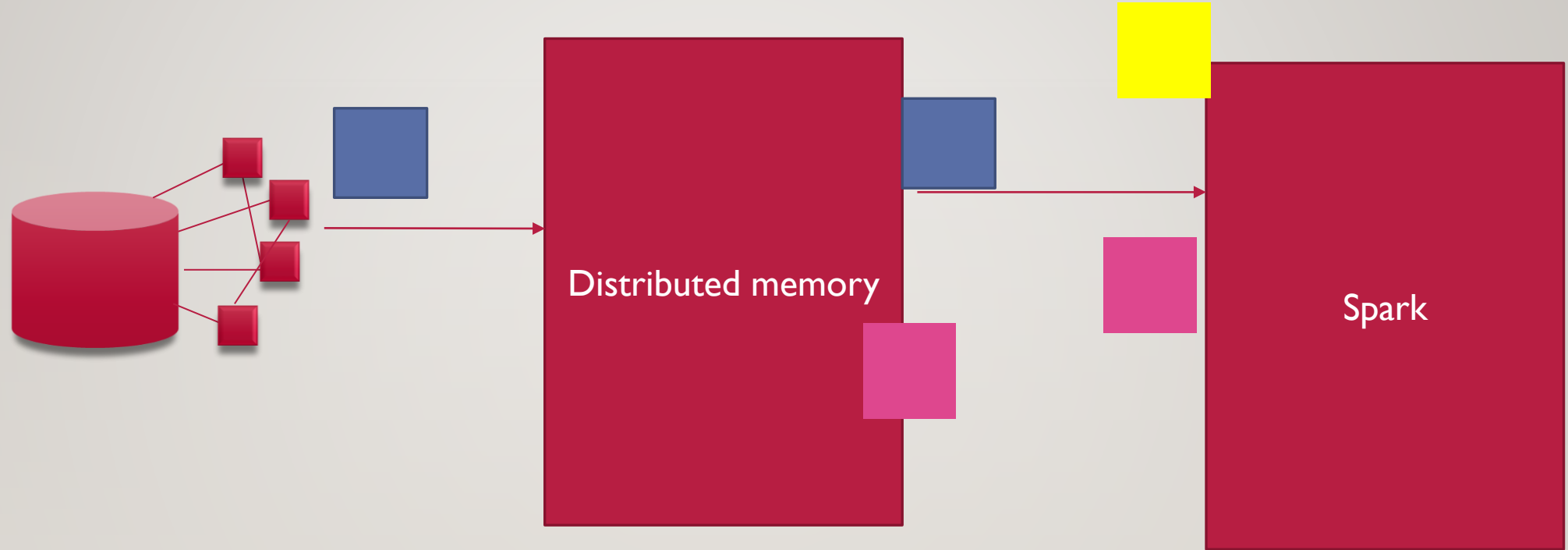
APACHE SPARK

- *Both batch processing and real time data processing*
- *100x faster in memory and 10x faster on disk since intermediate results not written to disk*
- *DAG has multiple vertices*

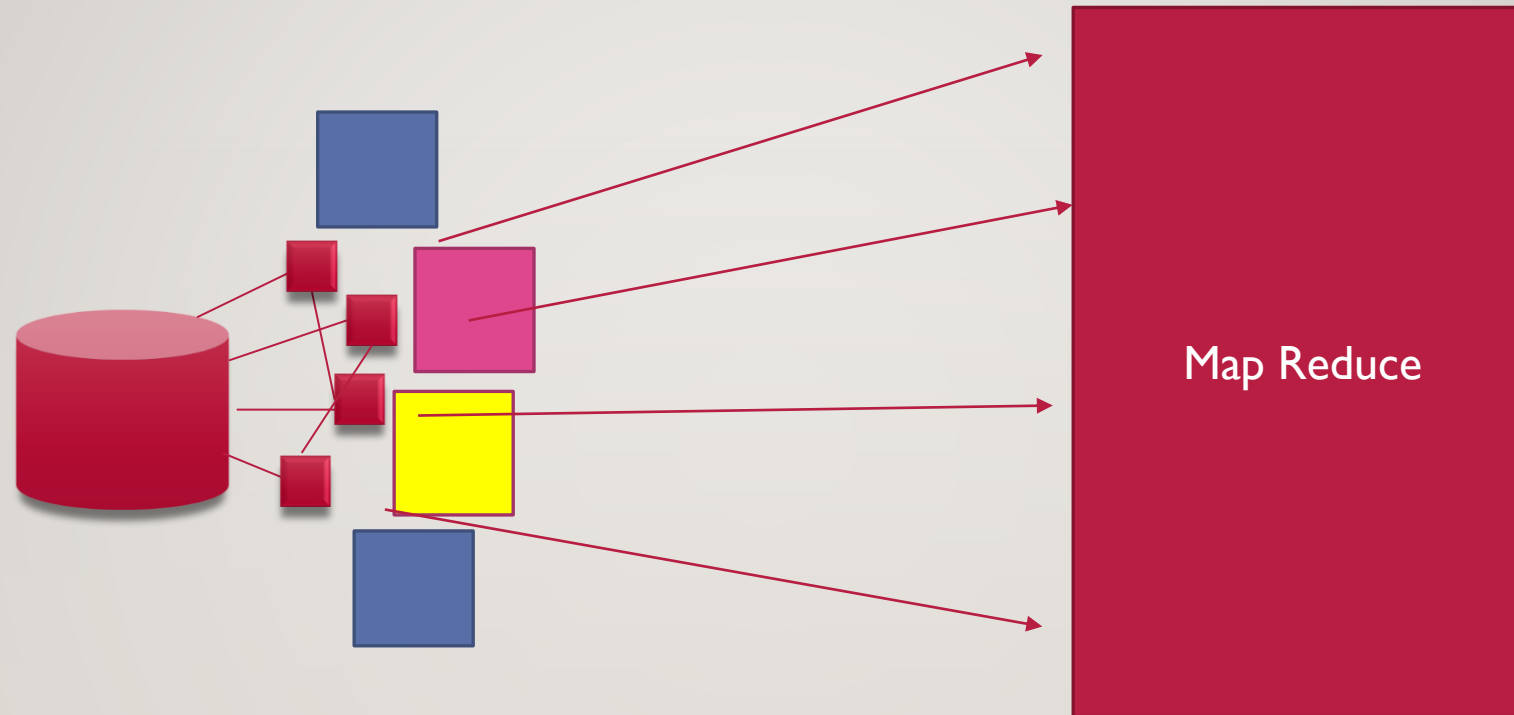
ITERATIVE PROCESSES



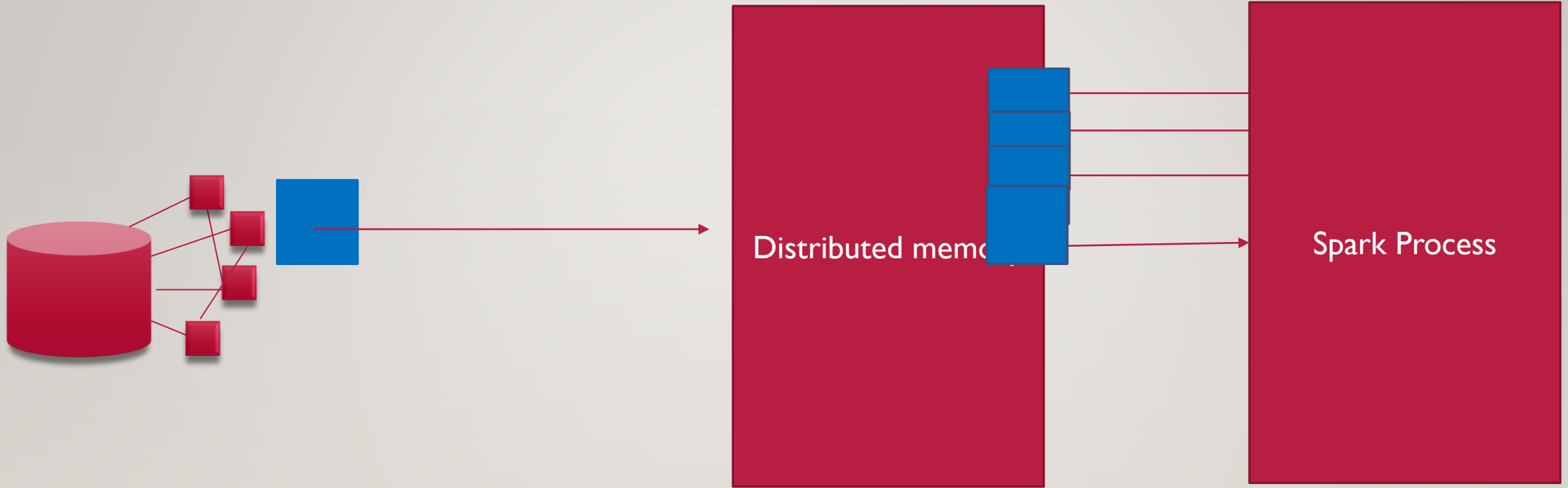
ITERATIVE PROCESSES



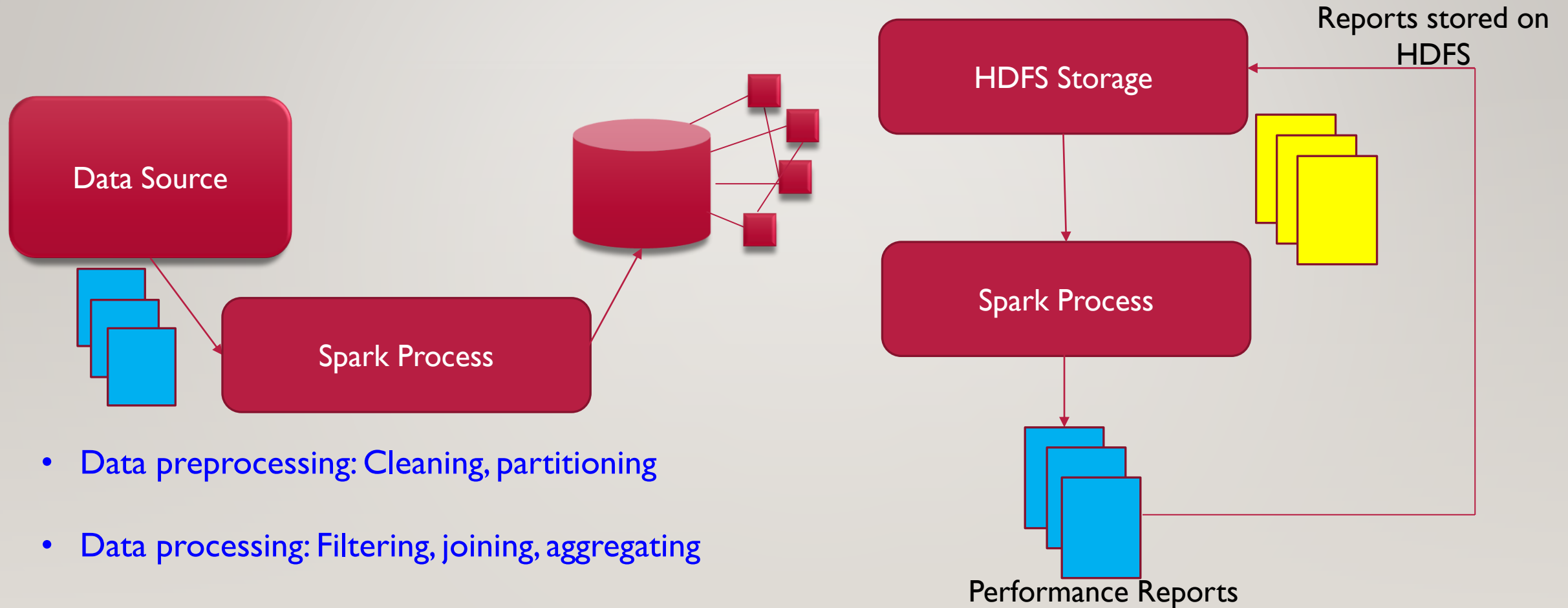
INTERACTIVE PROCESS(SQL QUERY)



INTERACTIVE PROCESS(SQL QUERY)



WHERE DO YOU SEE SPARK IN THE BIG DATA PIPELINE?

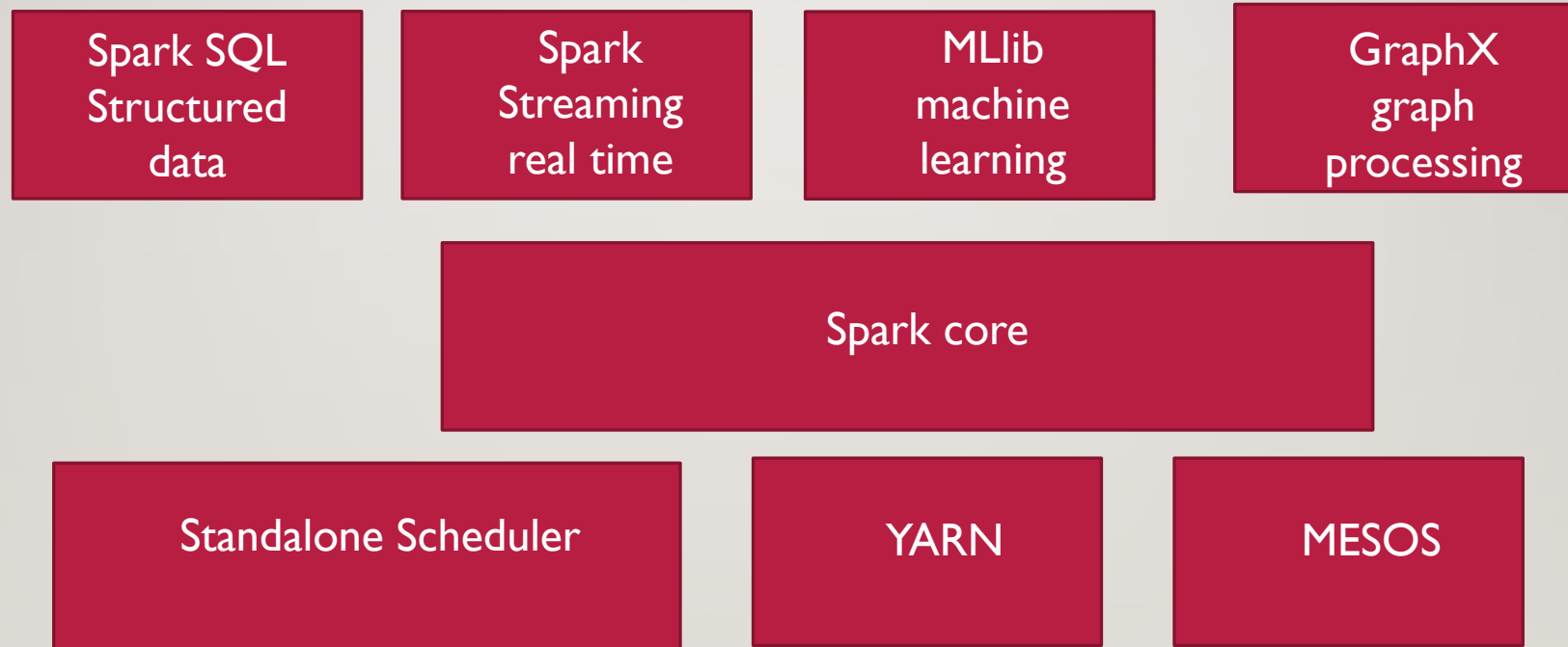


- Data preprocessing: Cleaning, partitioning
- Data processing: Filtering, joining, aggregating

ADVANTAGES OF SPARK

- Capable of handling several petabytes of data
- Simplicity
- Speed-100 times faster than Hadoop Map Reduce
- Spark is often used with distributed data stores such as MapR XD, Hadoop's HDFS, and Amazon's S3, with popular NoSQL databases such as MapR Database, Apache HBase, Apache Cassandra, and MongoDB, and with distributed messaging stores such as MapR Event Store and Apache Kafka.

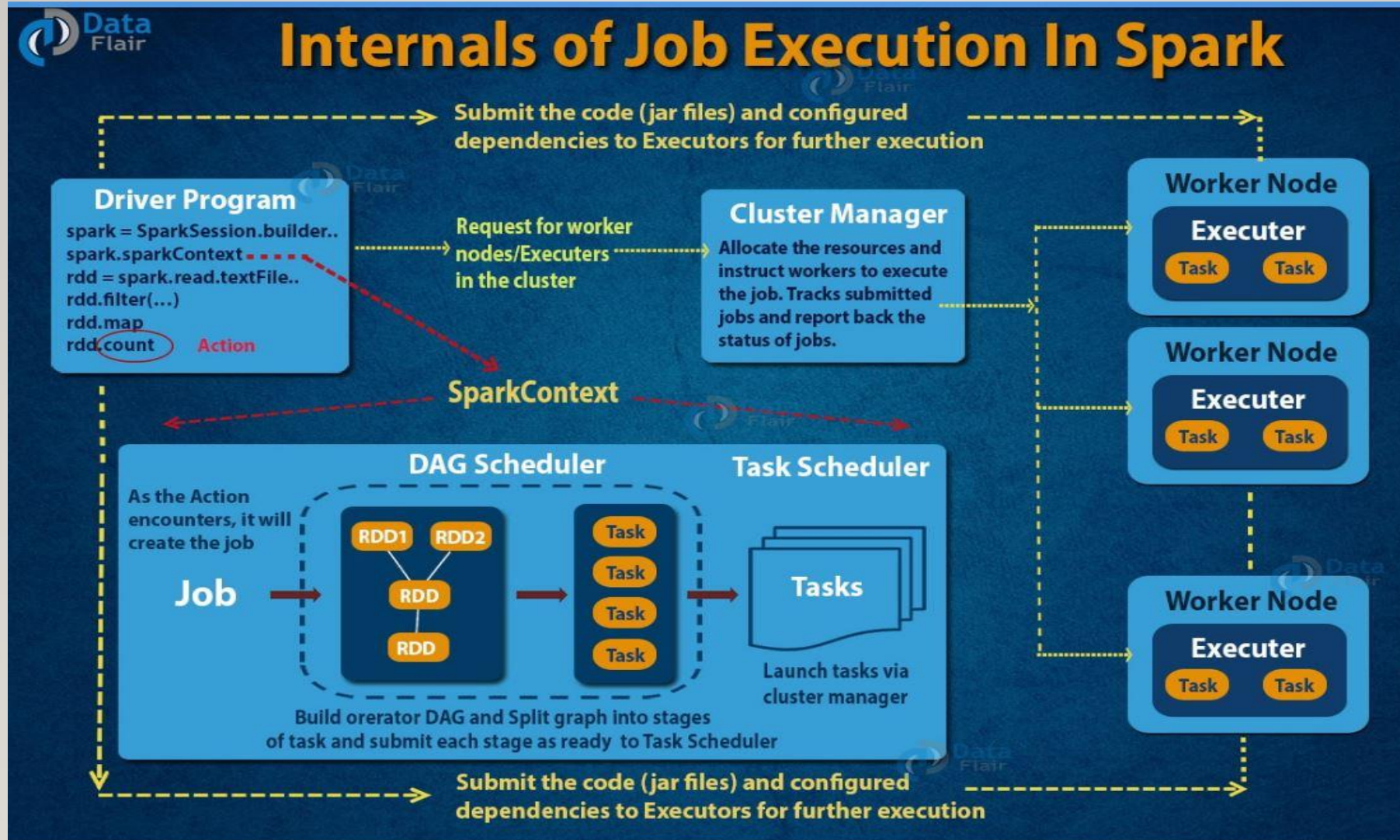
COMPONENTS OF SPARK



COMPONENTS OF SPARK

- All the functionalities of Apache Spark built on the top of **Spark Core**.
- Significant in programming and observing the role of the Spark Cluster
- **Spark SQL** is a Distributed framework for *structured* data processing and acts as a distributed SQL query engine.
- **Spark streaming** is an add-on to core Spark API which allows scalable, high-throughput, fault-tolerant stream processing of live data streams.
- **GraphX** in Spark is API for graphs and graph parallel execution.
- **MLlib** creation is to make machine learning scalable and easy.

HOW IT WORKS?



MODES OF OPERATION

Client Mode

Spark Driver and Spark
Context run on client
node outside Yarn
Cluster

Local mode

Process run in within a
node

Cluster mode

They run inside the
YARN cluster

USE CASES



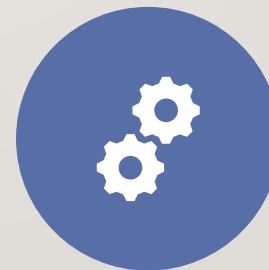
Stream processing



Machine learning



Interactive analytics



Data integration

WHO IS USING SPARK??

- IBM and Huawei have invested significant sums in the technology,
- The major Hadoop vendors, including MapR, Cloudera, and Hortonworks, have all moved to support YARN-based Spark alongside their existing products,
- Chinese search engine Baidu, e-commerce operation Taobao, and social networking company Tencent, all run Spark-based operations at scale, with Tencent's 800 million active users reportedly generating over 700 TB of data per day for processing on a cluster of more than 8,000 compute nodes.
- Pharmaceutical company Novartis depends upon Spark to reduce the time required to get modeling data into the hands of researchers
- Internet powerhouses such as Netflix, Yahoo, and eBay have deployed Spark at massive scale, collectively processing multiple petabytes of data on clusters of over 8,000 nodes. It has quickly become the largest open source community in big data, with over 1000 contributors from 250+ organizations.

DATA PREPROCESSING USING SPARK



Handling missing
values



Filtering of Garbage
data



Changing datatypes
and data with non-
conformities

WHAT ARE THE ISSUES THAT COME UP WHEN WE DO PRE-PROCESSING?

```
,code,c-value,segmentClosed,score,speed,average,reference,travelTimeMinutes,time,FID
```

```
0,1485581516,99,,30,63,64,64,0.41,2019-03-01 00:00:07 CST,34208
```

```
1,1485768829,98,,30,65,65,65,0.382,2019-03-01 00:00:07 CST,48672
```

```
Let's see if spark will take care of this one
```

```
5,
```

```
2,1485517139,97,,30,67,66,66,0.177,2019-03-01 00:00:07 CST,29294
```

```
3,1485706642,100,,30,65,65,65,0.675,1899-03-01 00:00:07 CST1899-03-01 00:00:07 CST1899-03-01 00:00:07 CST,44017
```

```
4,1485876925,,,10,61,61,61,0.56,2019-03-01 00:00:07 CST,55858
```

```
5,1485876676,99,,30,65,64,64,0.495,2019-03-01 00:00:07 CST,55835
```

```
6,1485613605,,,10,60,60,60,0.237,1899-03-01 00:00:07 CST,36855
```

```
7,1485621940,79,,30,71,65,65,0.3,2019-03-01 00:00:07 CST,37583
```

```
8,1485876574,98,,30,67,66,66,0.631,2019-03-01 00:00:07 CST,55826
```

Issue with
timestamps

Absence of
headers

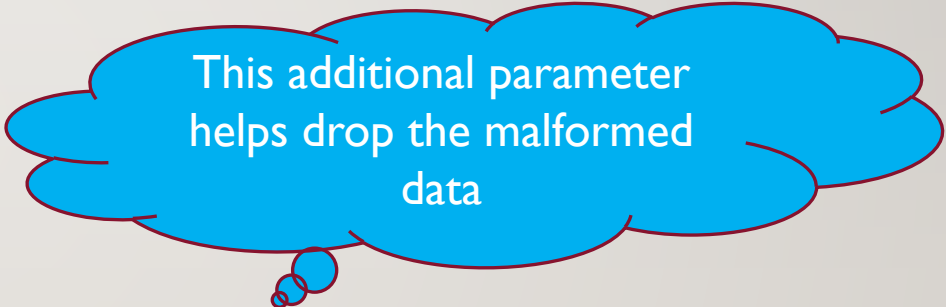
Malformed
data

ISSUES AND SOLUTIONS

Data issue	PySpark solution
Malformed data	An extra parameter included when reading data into spark data frame
Timestamp issue	The data can be dropped by first converting time string to timestamp and then filtering out based on year or regex pattern
Absence of headers	The headers and the datatypes of the columns can be specified in the initial step when defining the schema

HANDLING MALFORMED DATA

```
df1 = spark.read.csv('Data_cleaning6.csv',inferSchema=True, header =True)
```



This additional parameter
helps drop the malformed
data

```
df2=spark.read.format('csv').option("mode", "DROPMALFORMED").option('header', 'true')  
    .load('Data_cleaning6.csv',inferSchema=True)
```



Join datasets



Filter data based on certain conditions



Aggregate data by a column or group of columns and find average, sum etc

DATA POST- PROCESSING USING SPARK

RDD AND SPARK DATAFRAME

- Spark Data frame holds data in column and row format
- Column represents feature or variable
- A row represents an individual datapoint
- Spark started with RDD but Spark 2 and higher uses data frame syntax which is much easier to work with
- Dataframes used to apply transformations
- Results of the transformations can be display using shown or collected

Let's start coding....



PANDAS DATAFRAME

A 2- dimensional data structure aligned in a tabular fashion in rows and columns

Registration	Name	Marks
11010110	Steve	87
11010110	Jose	90
11010110	Patty	72
11010110	Vin	85

NUMPY ARRAY

- NumPy or Numerical Python helps with computing large sets of numerical data where python lists fail.
- In numpy arrays all the elements are of the same type and are fast and efficient
- Pandas is built on NumPy

IN A NUTSHELL

- Spark works on in memory processing and DAG while MapReduce involves more reading and writing from disk.
- Fast and efficient
- Spark is Open source and constantly evolving and there is growing interest and contribution
- It is a great option for transformation and analysis of big data (distributed operation)

REFERENCES

- <http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf>
- <https://data-flair.training/blogs/apache-spark-ecosystem-components/>
- <https://databricks.com/spark/>
- kdnuggets.com/2018/10/apache-spark-introduction-beginners.html