PROCESSING OF BIG DATA

SPARK

SESSION-1

FUN TIME

Introduction

- Apache Spark is an open source distributed data processing engine originated in UC Berkeley lab

- Project started in response to limitations of MapReduce framework

- Provides high level API for parallel data processing with inbuilt fault tolerant in distributed environment

Course outline

- Data holders: RDD & Dataframe/Dataset
- Spark SQL
- Supported file formats
- ML algorithms in Spark
- Graph processing using GraphX
- Job configuration parameters
- Optimization and performance tuning
- Spark streaming

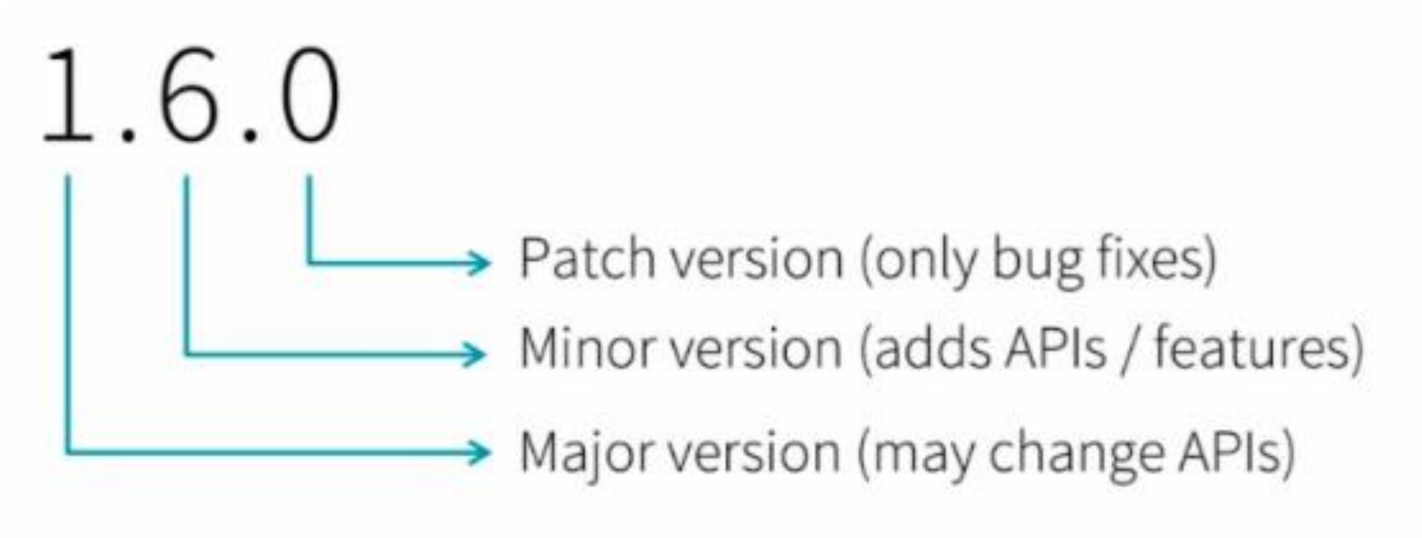
Agenda for today

- Install Spark on standalone mode
- RDD, Dataframe & Dataset
- RDD to Dataframe conversion
- Spark SQL

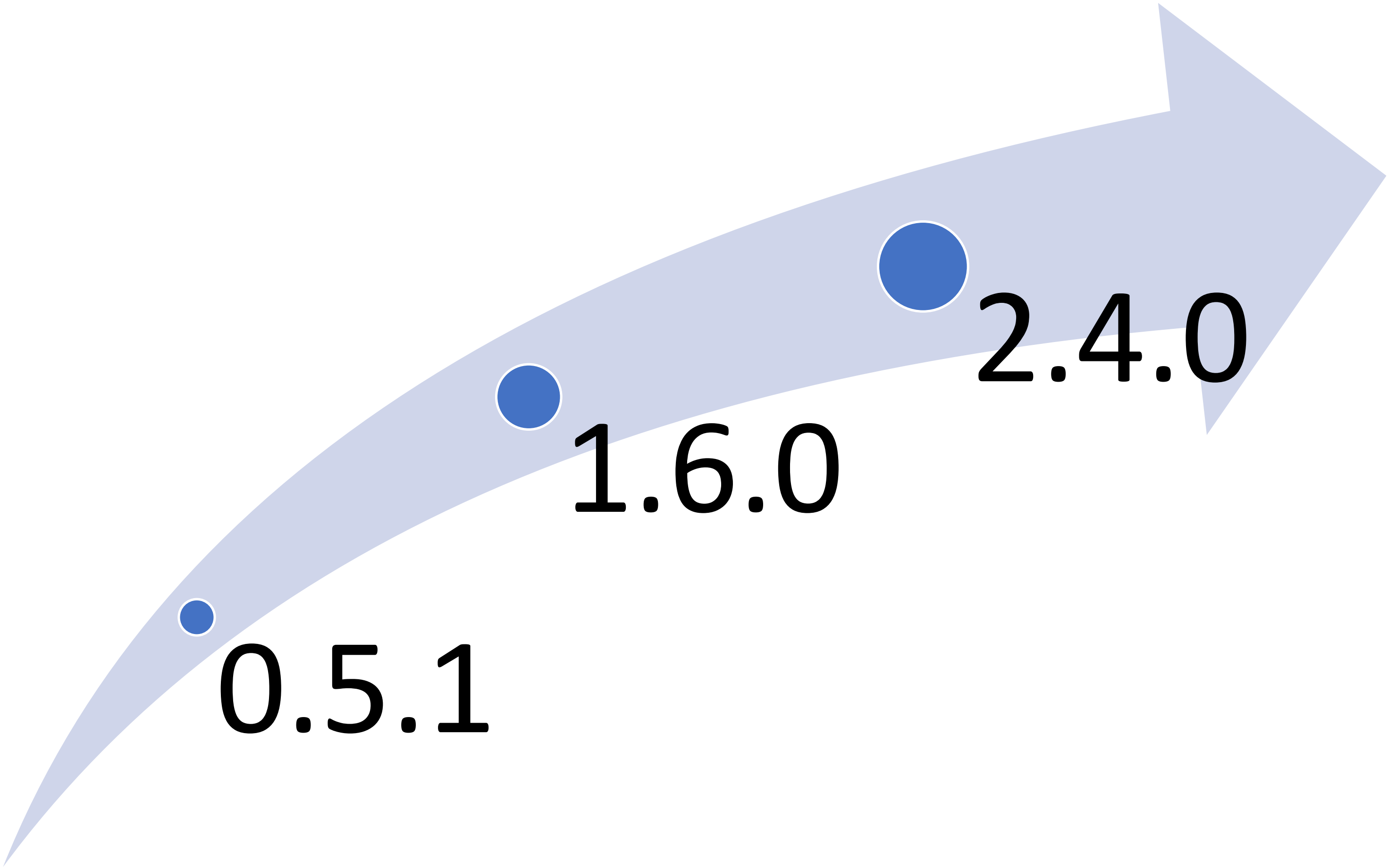
Warmup

- Copy a file from local to HDFS
- Read file from hdfs and display the count of records
- Display count of some filtered records
- Aggregate
- Save RDD in a persistent file system
- See the content of an HDFS file

How to interpret version#


1.6.0
Patch version (only bug fixes)
Minor version (adds APIs / features)
Major version (may change APIs)

Spark versions

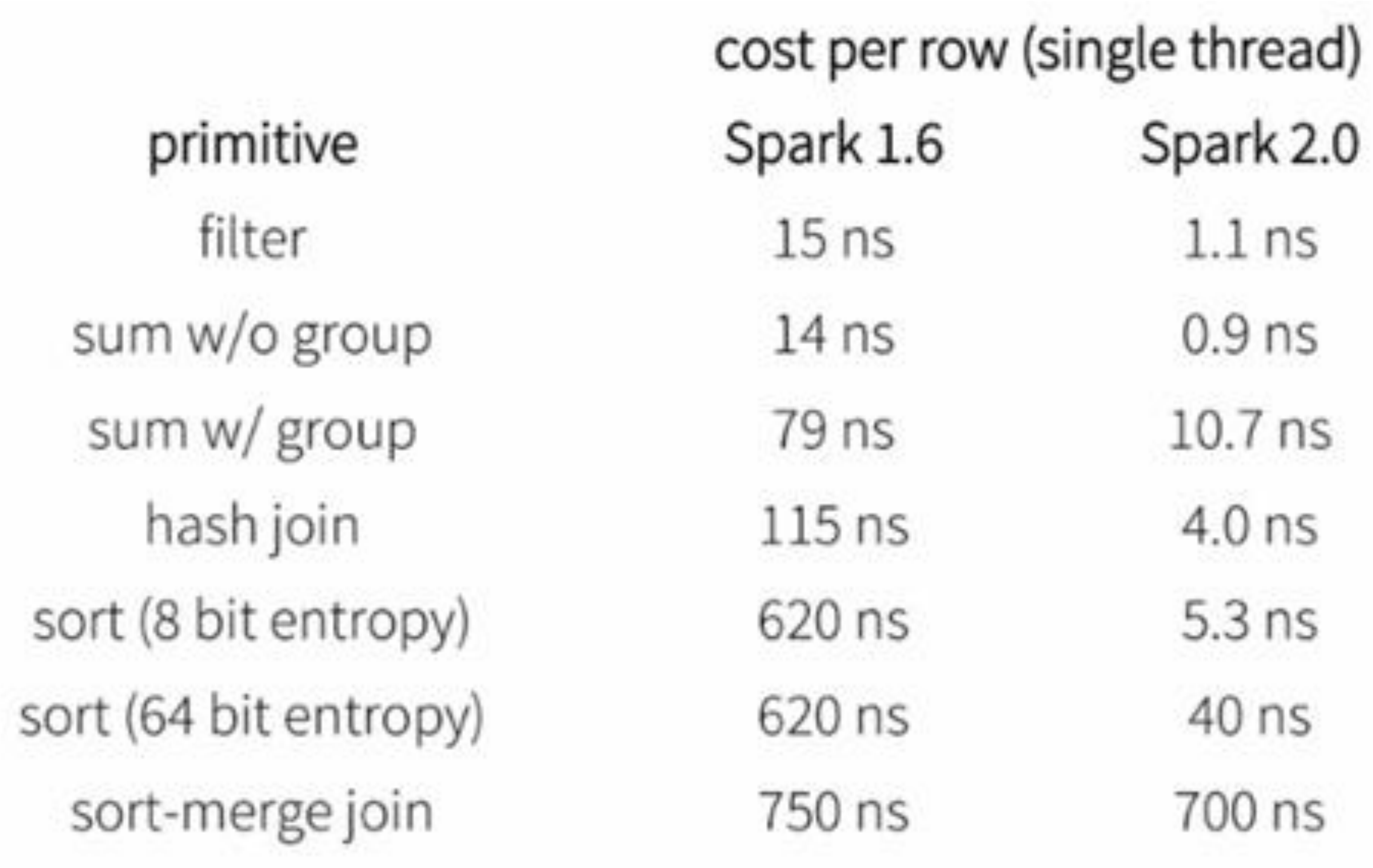


0.5.1

1.6.0

2.4.0

Perf. comparison



	cost per row (single thread)	
primitive	Spark 1.6	Spark 2.0
filter	15 ns	1.1 ns
sum w/o group	14 ns	0.9 ns
sum w/ group	79 ns	10.7 ns
hash join	115 ns	4.0 ns
sort (8 bit entropy)	620 ns	5.3 ns
sort (64 bit entropy)	620 ns	40 ns
sort-merge join	750 ns	700 ns

Install Spark 2.4.0 on Ubuntu

- Install python pip

  ```
  sudo apt-get install python-pip
  ```

- Install pyspark using pip

  ```
  sudo pip install pyspark
  ```

- Install JRE

  ```
  sudo apt install default-jre
  ```

- Export JAVA_HOME in .bashrc file
- Apply .bashrc file

  ```
  source ~/.bashrc
  ```

Reference

- https://spark.apache.org/docs/latest/sql-programming-guide.html

- https://community.cloud.databricks.com

- https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-python.html