Big Data (Data Ingestion & Data Pipeline)

To AGIT

September, 2019

Sharing Knowledge Requirement

Organizer:

- Server: Anaconda, Airflow, Airflow User
- Data: Open Data (CSV)

User:

- Laptop: IDE (Notepad++ / Sublime), MobaXterm, Chrome

What is Big Data?

Explosion of data and devices (IoT)

30B
connected devices

$200B
total market[1]

440x
more data

Transformation of IT infrastructure

big data

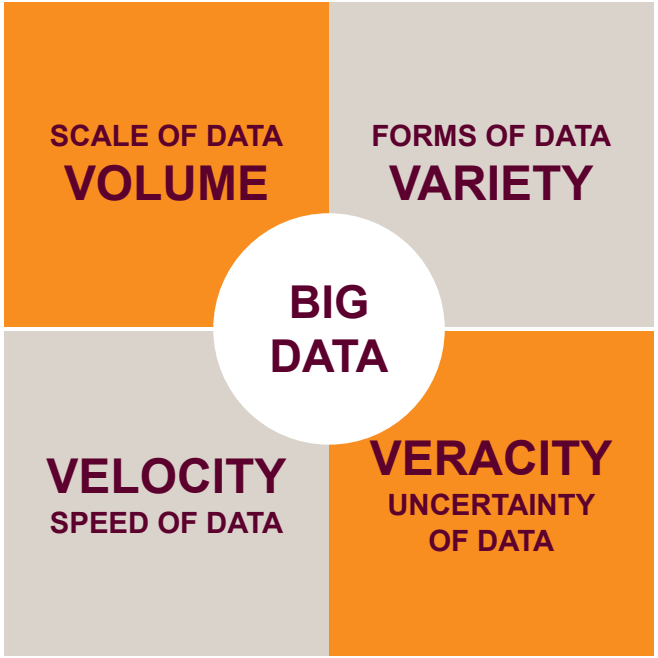
cloud

machine learning

The data-driven enterprise

[1] IDC **Worldwide Big Data and Business Analytics Market Through 2020**

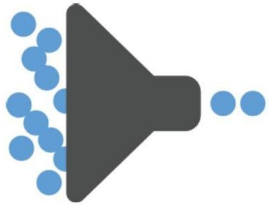
5V of Big Data



Big Data Analytic

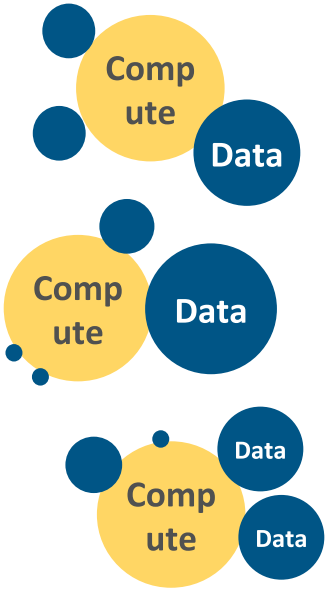
"Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights"

Data Pipeline



Paradigm Shift

LEGACY = Data to Compute

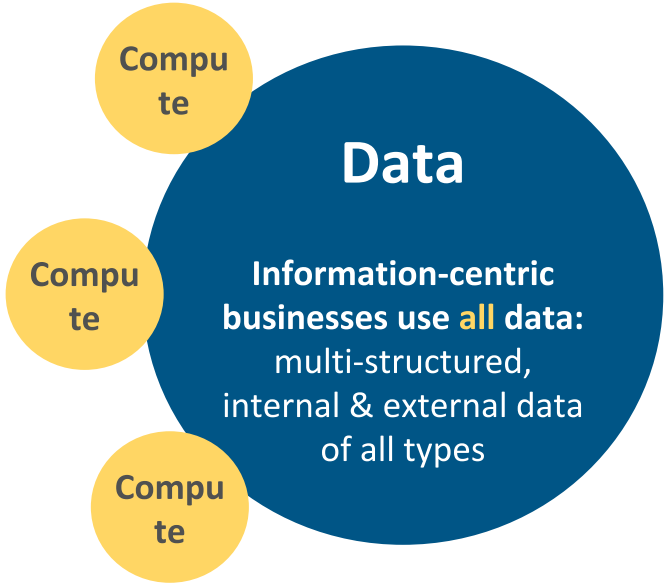


Process-centric businesses use:

Structured data mainly
Internal data only
“Important” data only

Siloed data sources

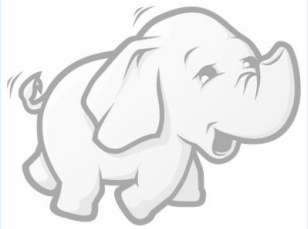
MODERN = Compute to Data



Hadoop

Hadoop is a platform for data storage and processing that is…

- ✔ Scalable
- ✔ Fault tolerant
- ✔ Open source



CORE HADOOP COMPONENTS

Hadoop Distributed File System (HDFS)

File Sharing & Data Protection Across Physical Servers

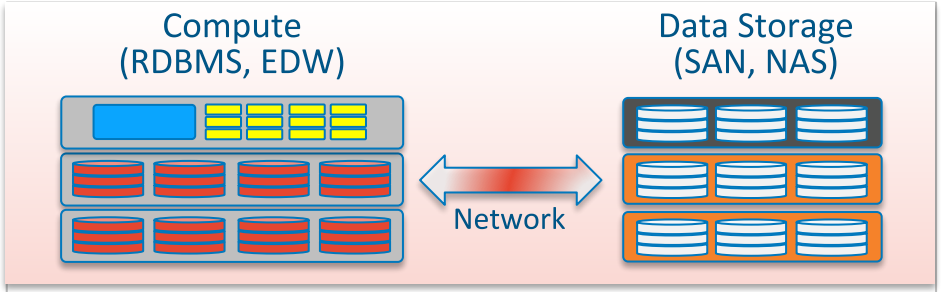
+

MapReduce

Distributed Computing Across Physical Servers

RDBMS vs Hadoop

The Old Way

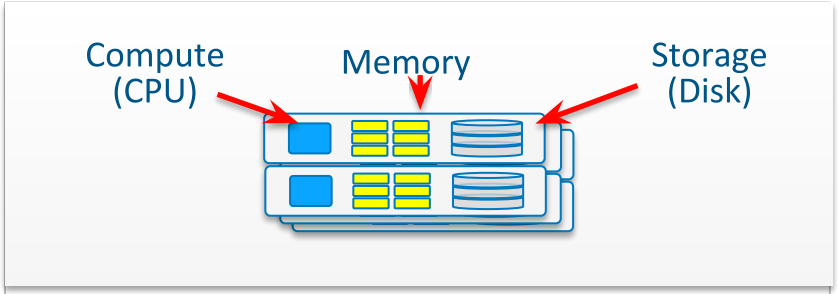


Expensive, Special purpose, "Reliable" Servers
Expensive Licensed Software

Expensive & Unattainable

$30,000+ per TB

The Hadoop Way



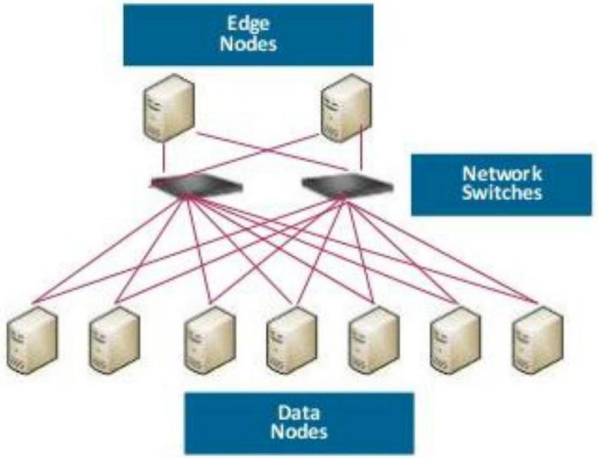
Industry Standard Servers
Hybrid Open Source Software

Affordable & Attainable

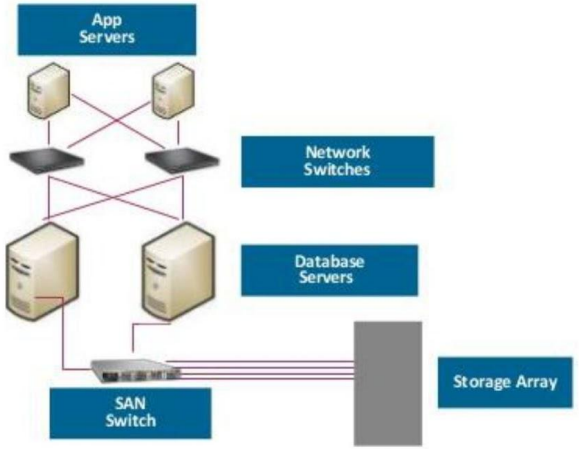
$300-$1,000 per TB

Hadoop Infrastructure

Hadoop



Data Warehouse/RDBMS

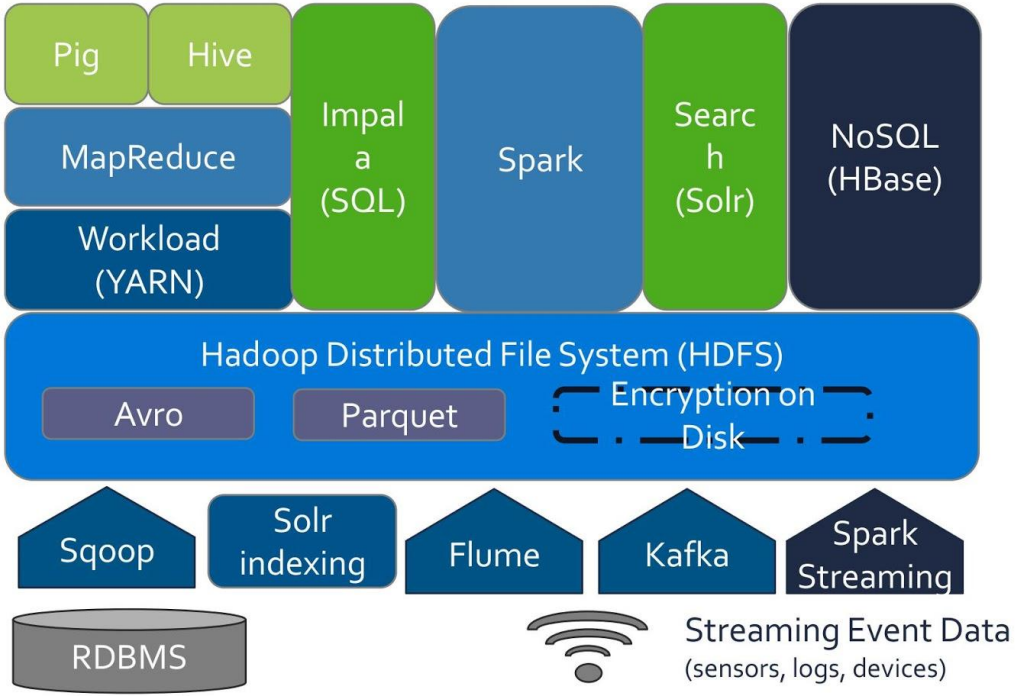


Hadoop Solution Provider

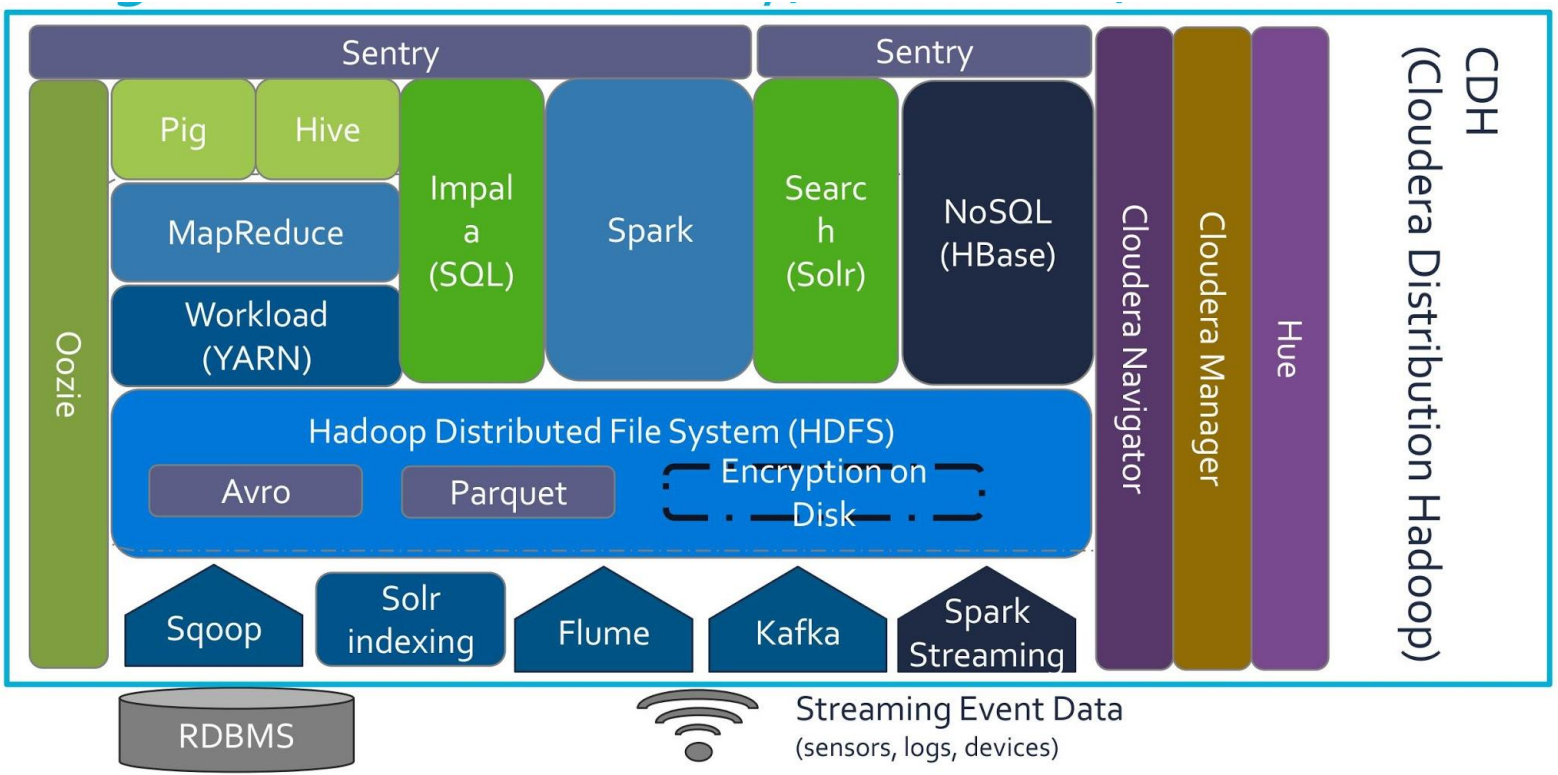
cloudera®

MAPR®

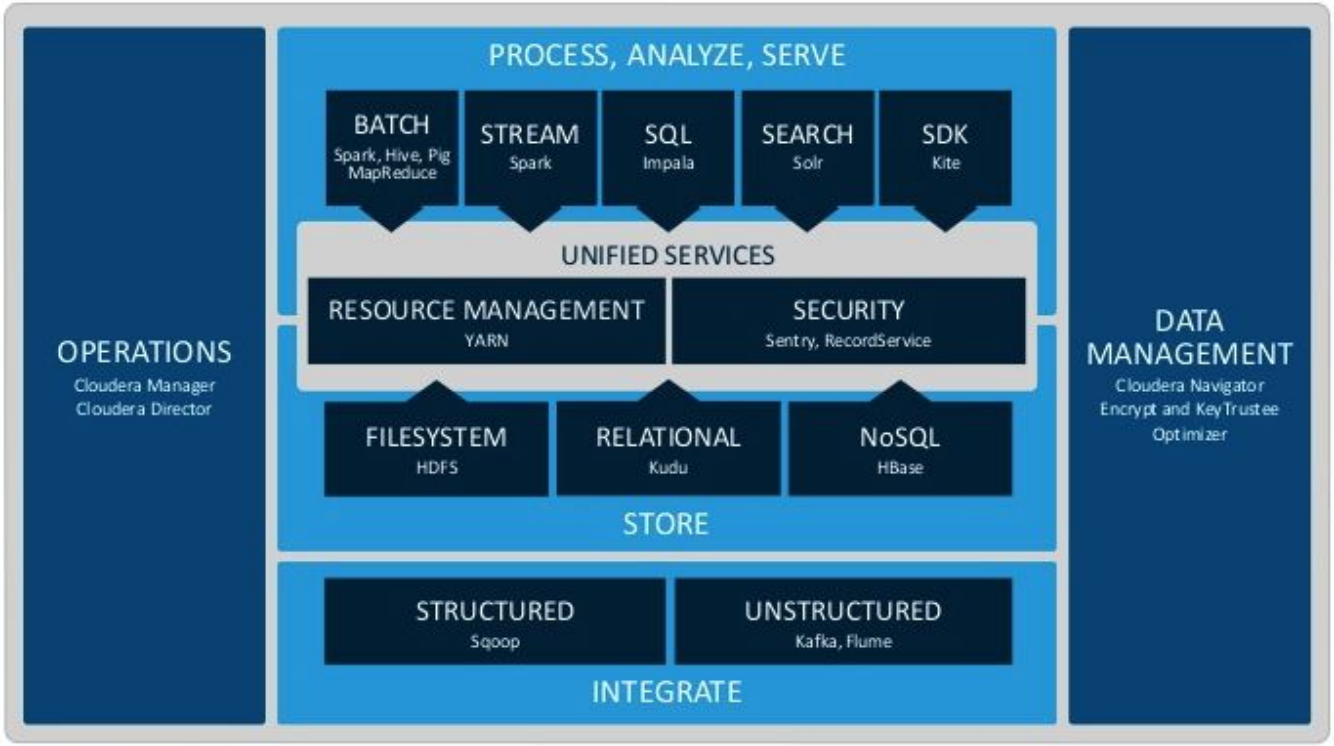
Hadoop Environment



Cloudera Distribution Hadoop (CDH)



Cloudera Services



Cloudera Customer

Financial Services

Telco

Healthcare & Life Sciences

Media & Technology

Retail & CP

Public Sector

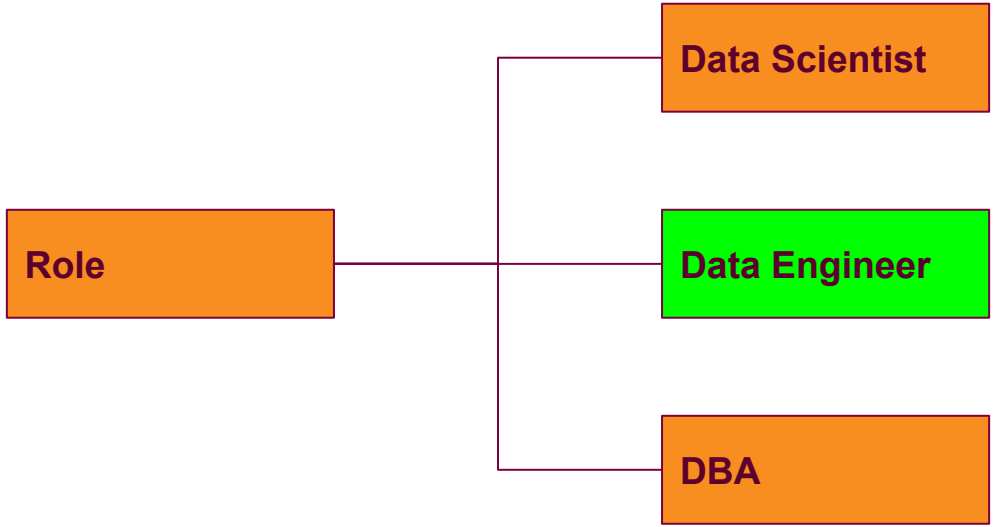
Big Data Impact

HelloFresh: updating 2500 BI dashboards daily for data-driven decisions

Zurich Insurance: using data insights to deliver personalised services, custom policies and reduce risk

Discussion

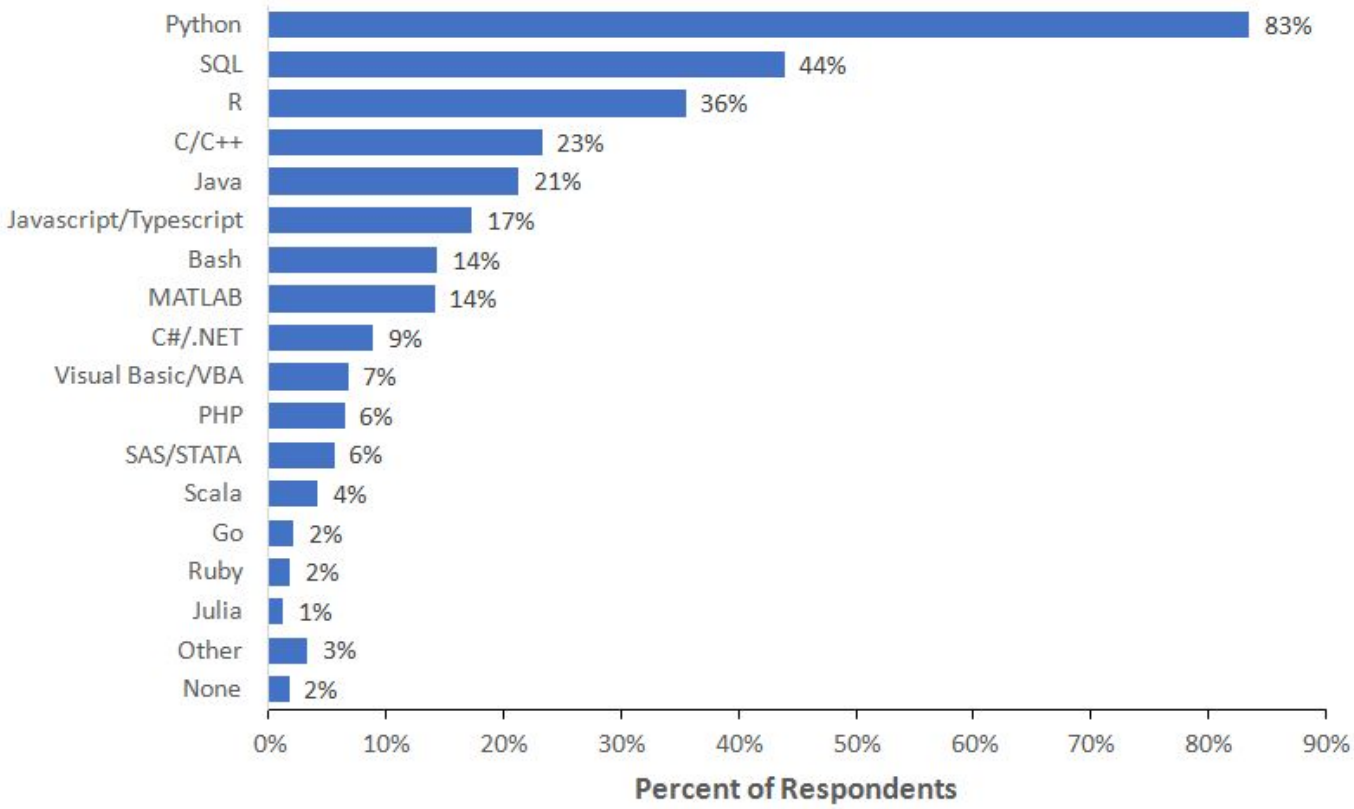
Big Data Role



Why Python?

Programming Language

Data Programming Language



Open Notebook
(On Google Colab)

Variable and Types

Strings

“Ini adalah String”

Numbers

120
123.45

Open Notebook

Array and Dictionary

Array

[1,2,3,4]

Dict

{
“A” : 1
“B” : 2
“C” : 3
}

Open Notebook

Operator

Strings

String Formatting

Number

+
-
:
*****
%

Open Notebook

Exercise Time!

Loop

For

```
numbers = [1, 2, 3]
for number in numbers:
    print(number)
```

While

```
count = 0
while count < 5:
    print(count)
    count += 1
```

Open Notebook

Condition

If

```
name = "John"
if name in ["John", "Rick"]:
    print("Your name is either John or Rick.")
```

Open Notebook

Exercise Time!

Method

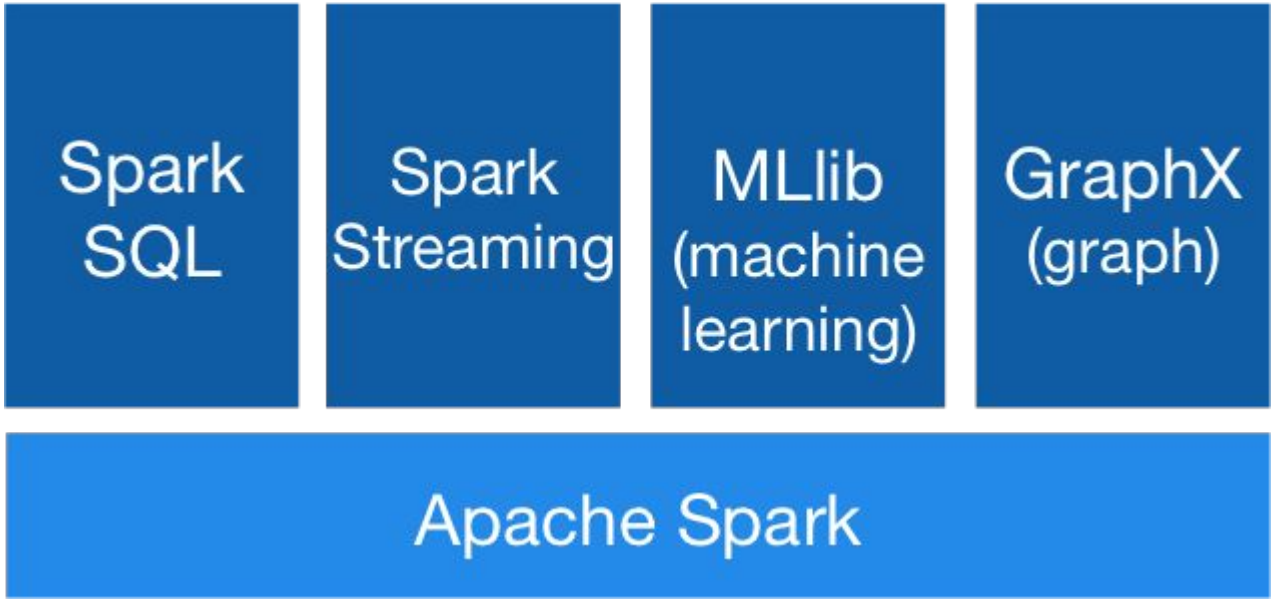
```
def checkNone(value1 = None, value2 = None):
  if value1:
    if value2:
      print("There's 2 value: ({},{})".format(value1,value2))
    else:
      print("There's 1 value: ({})".format(value1))
  elif value2:
    print("There's 1 value: ({})".format(value2))
  else:
    print("There's no value")

checkNone(value1 = 2, value2 = 3)
```

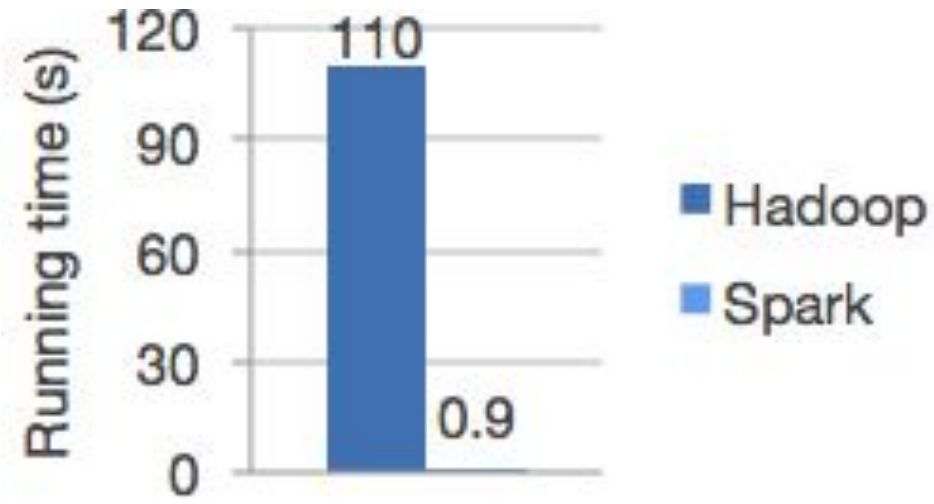
Open Notebook

Why Spark?

Spark: Jack of All Trades



Spark vs MapReduce



Open Notebook
(On Google Colab)

Initiate Spark

SparkSession

[SparkContext, HiveContext]

SparkContext

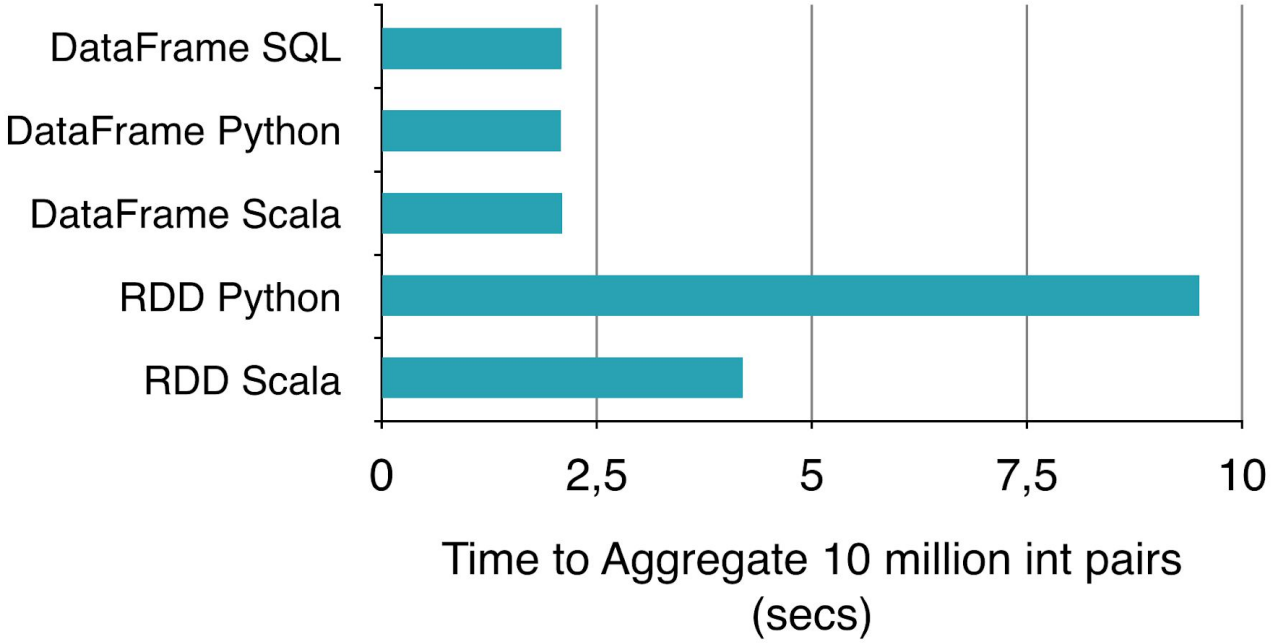
SparkContext

Open Notebook

Resilient Distributed Dataset (RDD)

Open Notebook

Physical Execution: Unified Across Languages



DataFrames

Tabular

Distributed

Analytic

Open Notebook

Exercise Time!

Why Airflow?

Oozie vs Airflow

- XML
- Restricted
- Small User

- Python
- Dynamic
- Industry Standard

Airflow

Dynamic

Extensible

Elegant

Scalable

Airflow List DAG

Airflow | DAGs | Data Profiling | Browse | Admin | Docs | About | 2018-09-07 22:14:10 UTC

DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	On	example_bash_operator	0 0 * * *	airflow	6	2018-09-06 00:00	5	
	On	example_branch_dop_operator_v3	*/1 * * * *	airflow	3 1 1 5	2018-09-05 00:56	54 3	
	On	example_branch_operator	@daily	airflow	5	2018-09-06 00:00	2	
	On	example_xcom	@once	airflow	3	2018-09-05 00:00	1	
	On	latest_only	4:00:00	Airflow	2	2018-09-07 16:00	35	

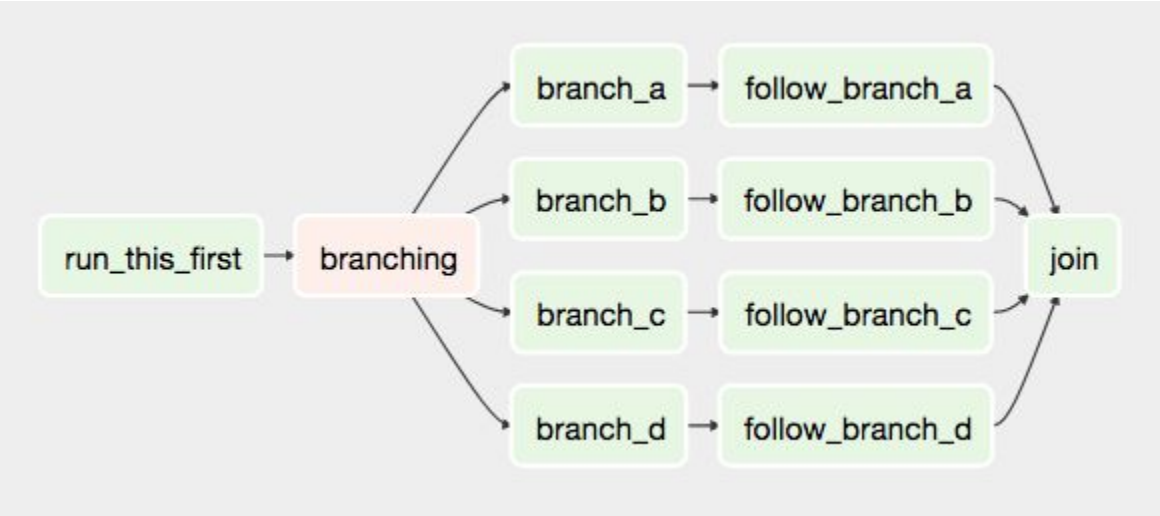
Showing 1 to 5 of 5 entries

« < 1 > »

Show Paused DAGs

Open Airflow

Airflow DAG



Open Airflow

Airflow Code Structure

- Importing Modules
- Default Arguments
- Instantiate a DAG
- Tasks
- Templating with Jinja
- Setting up Dependencies

Open Airflow

Exercise Time!

thank you ☺