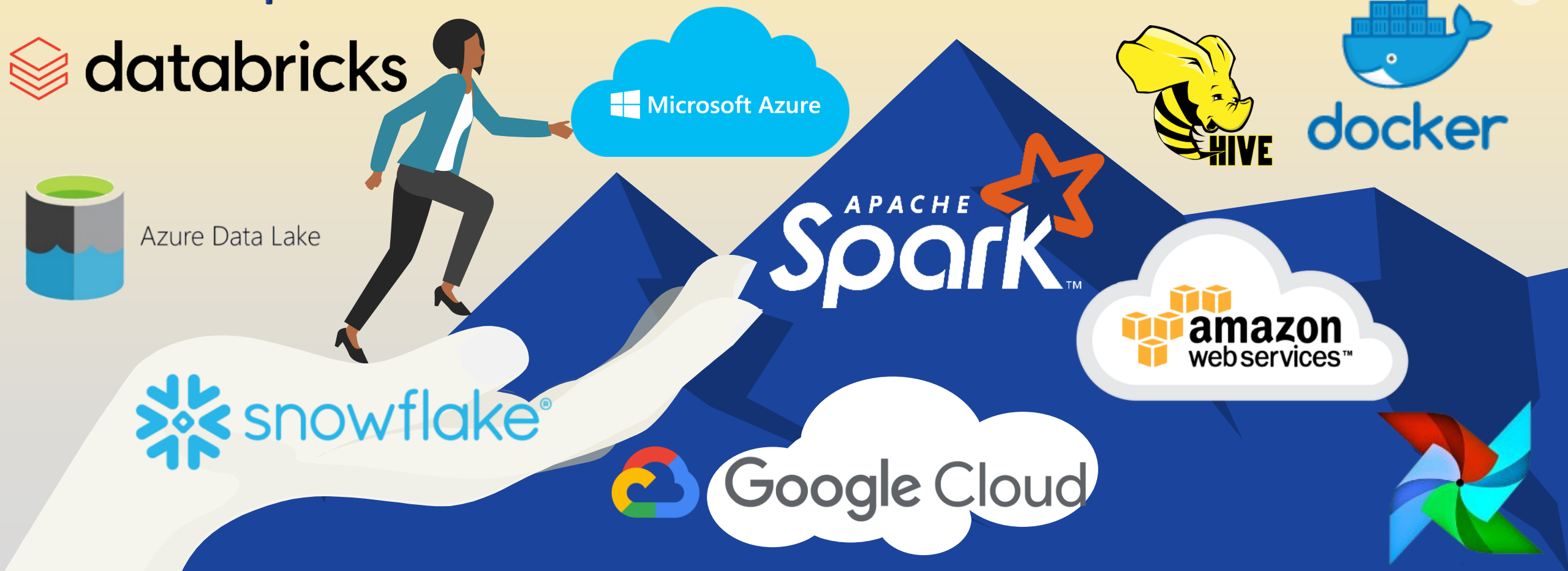
BIGDATAPEDIA

Learning will lead the future...

Full Stack Data Engineer

Experience with Multi Node Cluster & Cloud Services



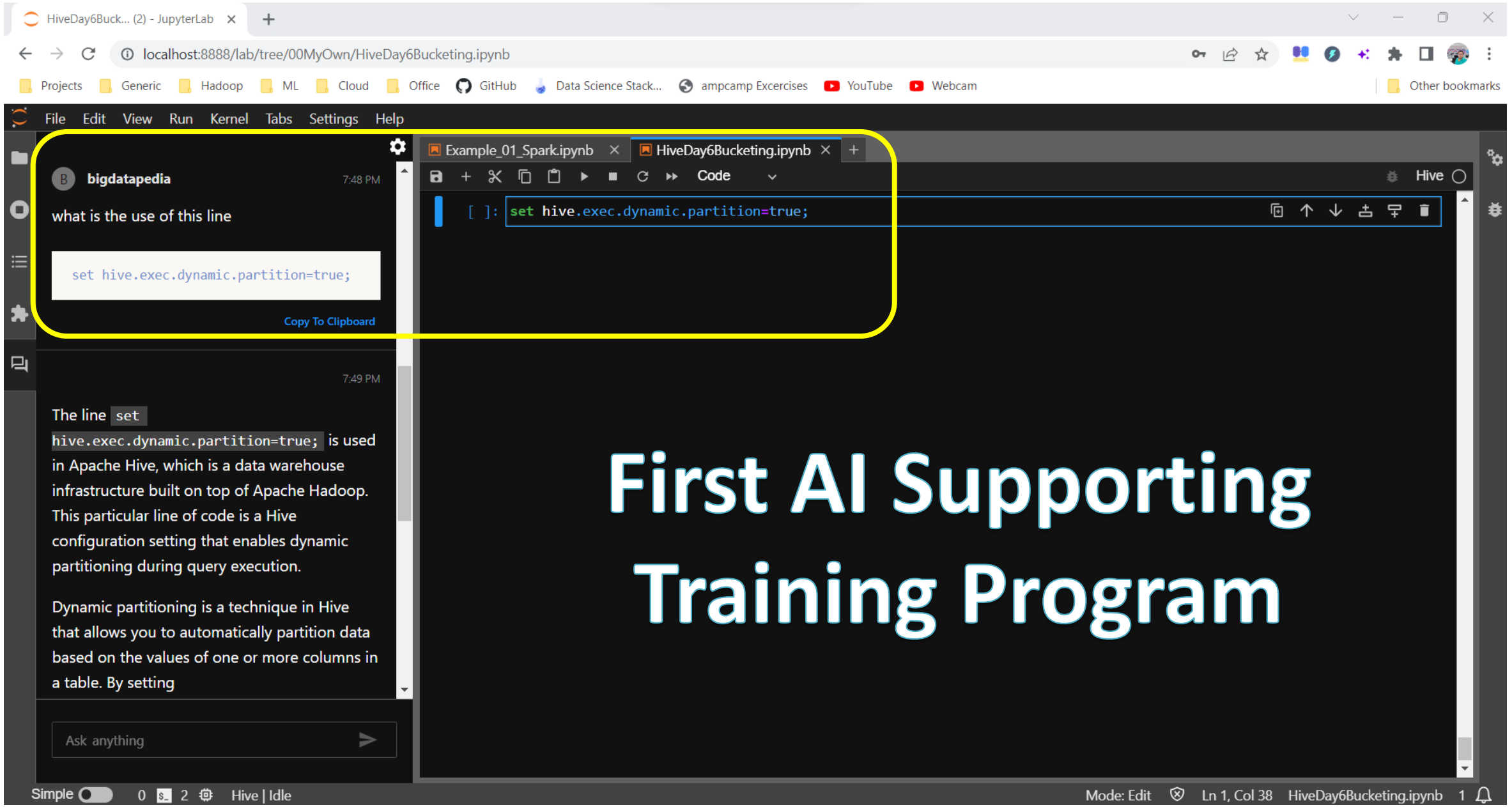
Experience with Multi Node Cluster

3-5 Multi Node Cluster

Presented By

BigDatapedia ML

Contact us @ +91 9715 010 010


HiveDay6Buck... (2) - JupyterLab
localhost:8888/lab/tree/00MyOwn/HiveDay6Bucketing.ipynb
Projects
Generic
Hadoop
ML
Cloud
Office
GitHub
Data Science Stack...
ampcamp Excercises
YouTube
Webcam
Other bookmarks
File
Edit
View
Run
Kernel
Tabs
Settings
Help
bigdatapedia
7:48 PM
what is the use of this line
set hive.exec.dynamic.partition=true;
Copy To Clipboard
7:49 PM
The line set hive.exec.dynamic.partition=true; is used in Apache Hive, which is a data warehouse infrastructure built on top of Apache Hadoop. This particular line of code is a Hive configuration setting that enables dynamic partitioning during query execution.
Dynamic partitioning is a technique in Hive that allows you to automatically partition data based on the values of one or more columns in a table. By setting
Ask anything
Example_01_Spark.ipynb
HiveDay6Bucketing.ipynb
Code
Hive
[ ]: set hive.exec.dynamic.partition=true;
First AI Supporting Training Program
Simple
0
2
Hive | Idle
Mode: Edit
Ln 1, Col 38
HiveDay6Bucketing.ipynb
1

All Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources
3	0	1	2	2	<memory:1792, vCores:2>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unl
5	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum
Capacity Scheduler	[<name=memory-mb default-unit=Mi type=COUNTABLE>, <name=vcores default-unit= type=COUNTABLE>]	<memory:256, vC

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB
application_1638167425795_0003	root	insert into newbatch values (1,'a','1234') (Stage-1)	MAPREDUCE	dev	0	Mon Nov 29 12:13:30 +0550 2021	Mon Nov 29 12:13:30 +0550 2021	N/A	RUNNING	UNDEFINED	2	2	1792
application_1638167425795_0002	root	PySparkShell	SPARK	prod	0	Mon Nov 29 12:04:46 +0550 2021	Mon Nov 29 12:04:47 +0550 2021	Mon Nov 29 12:06:00 +0550 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A
application_1638167425795_0001	root	PySparkShell	SPARK	prod	0	Mon Nov 29 12:03:32 +0550 2021	Mon Nov 29 12:03:33 +0550 2021	Mon Nov 29 12:04:13 +0550 2021	FINISHED	SUCCEEDED	N/A	N/A	N/A

Showing 1 to 3 of 3 entries

Multi Node Cluster In Realtime Practice

Windows PowerShell

```
                ******************************************
                *          Welcome to BigDatapedia          *
                ******************************************

Note* If you find any issues on this images please raise it in GitHub page or reach us via WhatsApp +91 9715 010 010
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia# pyspark --master yarn --deploy-mode client
Python 3.9.2 (default, Feb 28 2021, 17:03:44)
[GCC 10.2.1 20210110] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/11/29 06:34:44 WARN conf.HiveConf: HiveConf of name hive.metastore.event.db.notification.api.auth does not exist
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.2
      /_/

Using Python version 3.9.2 (default, Feb 28 2021 17:03:44)
SparkSession available as 'spark'.
>>> sc
<SparkContext master=yarn appName=PySparkShell>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```

Spark YARN Realtime Deployment

Summary

Security is off.

Safemode is off.

258 files and directories, 246 blocks = 504 total filesystem object(s).

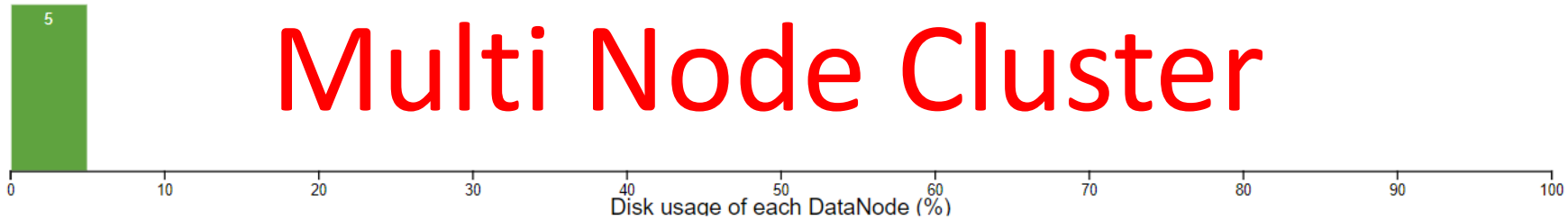
Heap Memory used 102.64 MB of 213.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 53.78 MB of 54.78 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	1.23 TB
DFS Used:	636.58 MB (0.05%)
Non DFS Used:	164.78 GB
DFS Remaining:	1 TB (81.71%)
Block Pool Used:	636.58 MB (0.05%)
DataNodes usages% (Min/Median/Max/stdDev):	0.04% / 0.05% / 0.06% / 0.01%
Live Nodes	5 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0

Multi Node Cluster

Datanode usage histogram



In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
73ff37c22de1:50010 (192.168.112.11:50010)	http://73ff37c22de1:50075	0s	0m	250.98 GB	152	125.26 MB (0.05%)	2.10.1
bc94daad9258:50010 (192.168.112.6:50010)	http://bc94daad9258:50075	0s	0m	250.98 GB	149	105.96 MB (0.04%)	2.10.1
cc43bc7caef2:50010 (192.168.112.3:50010)	http://cc43bc7caef2:50075	0s	1m	250.98 GB	149	146.49 MB (0.06%)	2.10.1
f8f9e903550d:50010 (192.168.112.4:50010)	http://f8f9e903550d:50075	0s	0m	250.98 GB	141	126.88 MB (0.05%)	2.10.1
fcb279b7c7ec:50010 (192.168.112.13:50010)	http://fcb279b7c7ec:50075	0s	0m	250.98 GB	147	131.99 MB (0.05%)	2.10.1

Windows PowerShell

```
                ******************************************
                *          Welcome to BigDatapedia          *
                ******************************************

Note* If you find any issues on this images please raise it in GitHub page or reach us via WhatsApp +91 9715 010 010
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia#
root@f7503dd58916:/home/bigdatapedia# hive
/home/bigdatapedia/hive/bin/hive: line 354: ps: command not found
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/bigdatapedia/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/bigdatapedia/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/home/bigdatapedia/hive/lib/hive-common-2.3.7.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using H
ive 1.X releases.
hive>
    > show databases;
OK
default
demo
Time taken: 0.641 seconds, Fetched: 2 row(s)
hive>
    > use demo;
OK
Time taken: 0.024 seconds
hive>
    > create external table newbatch (id int, name string, phno string)
    >  row format delimited fields terminated by ','
    > stored as textfile
    > location '/user/bigdatapedia/newbatch';
OK
Time taken: 0.194 seconds
hive> show tables;
OK
newbatch
```

BIGDATAPEDIA
Learning will lead the future...

Course Overview

- Topic 1 – BigData Ecosystem
- Topic 2 – Azure Data Engineering
- Topic 3 – AWS Data Engineering
- Topic 4 – Google Cloud Platform
- Topic 5 – Snowflake
- Topic 6 – Databricks

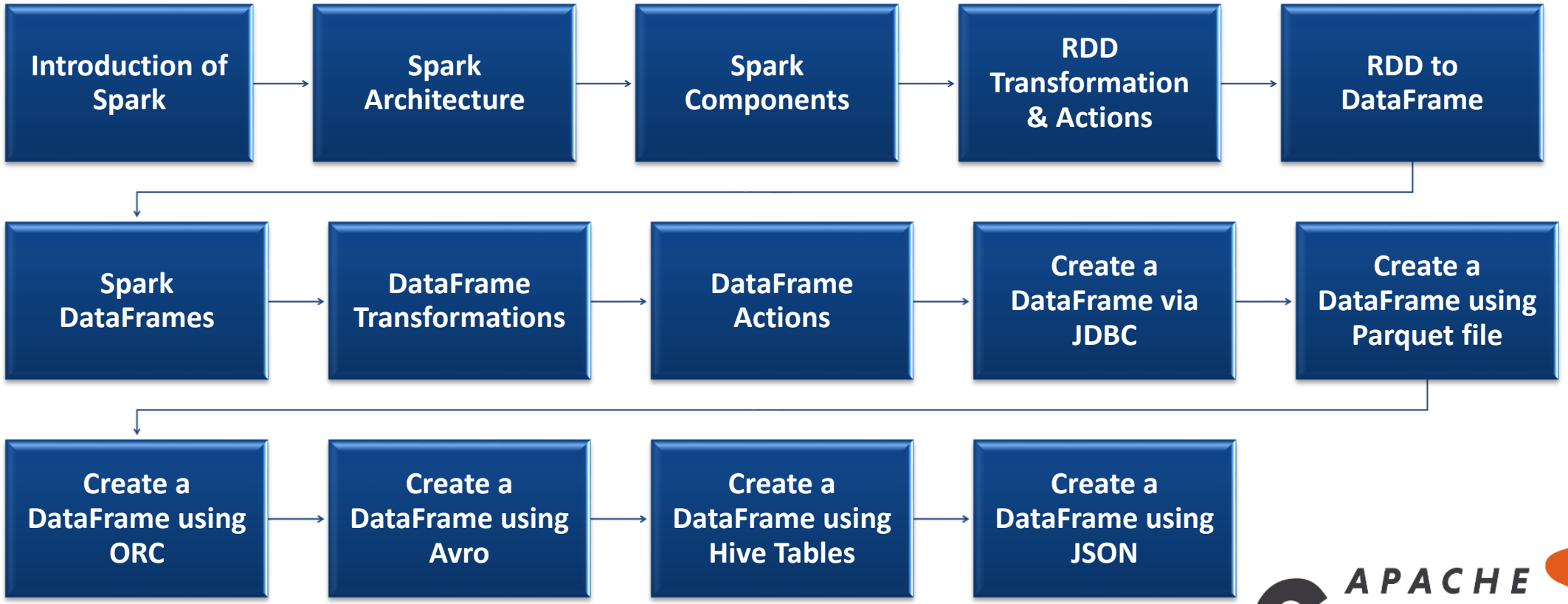
BIGDATAPEDIA
Learning will lead the future...

BigData Ecosystem
Hive Introduction

- Introduction on Hive
- Hive Architecture
- Hive Meta Store
- Hive Server 1 vs 2
- Beeline VS Hive CLI
- Connect hive via Beeline
- Thrift Server
- Managed VS External Table
- Create Table using Avro
- Create Table with Parquet
- Create Table with ORC
- Partitioning
- Partition Types
- Bucketing
- SerDe
- Hive Joins
- Complex Data Types (JSON)
- Top 10 Hive Performance Tuning

BIGDATAPEDIA
Learning will lead the future...

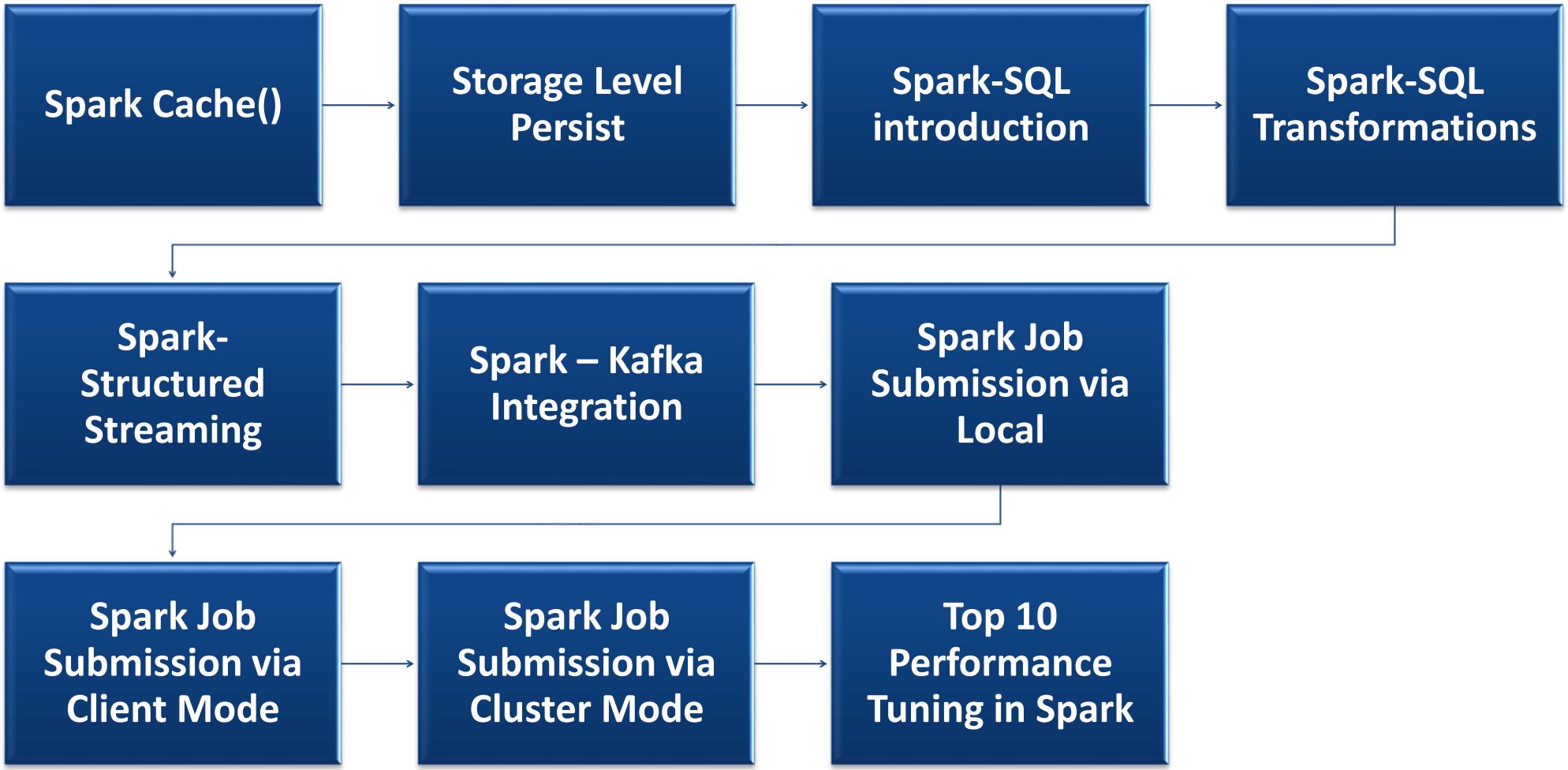
BigData Ecosystem
Apache Spark Introduction



continue...

BIGDATAPEDIA
Learning will lead the future...

BigData Ecosystem
Apache Spark Introduction (cont..)

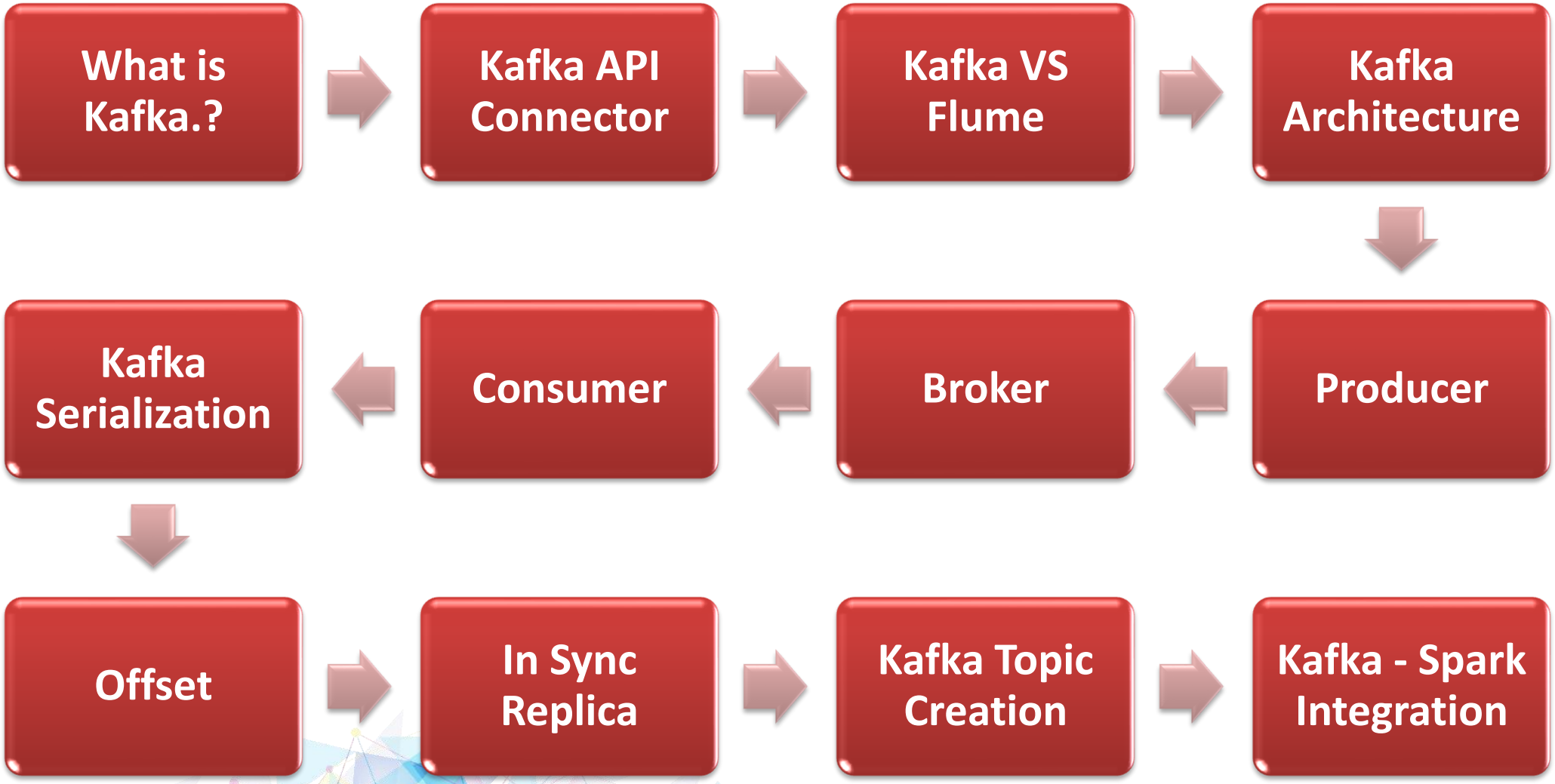


APACHE Spark™

BIGDATAPEDIA
Learning will lead the future...

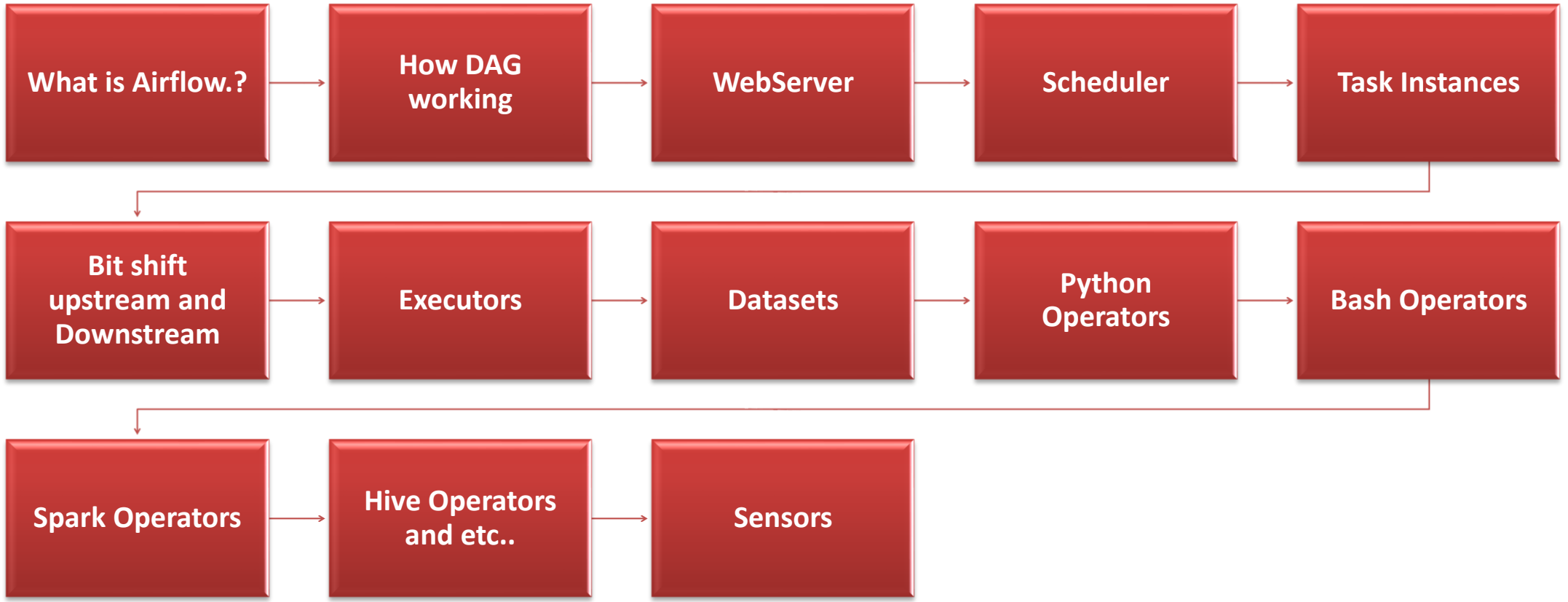
BigData Ecosystem
Kafka Introduction

APACHE kafka®
A distributed streaming platform



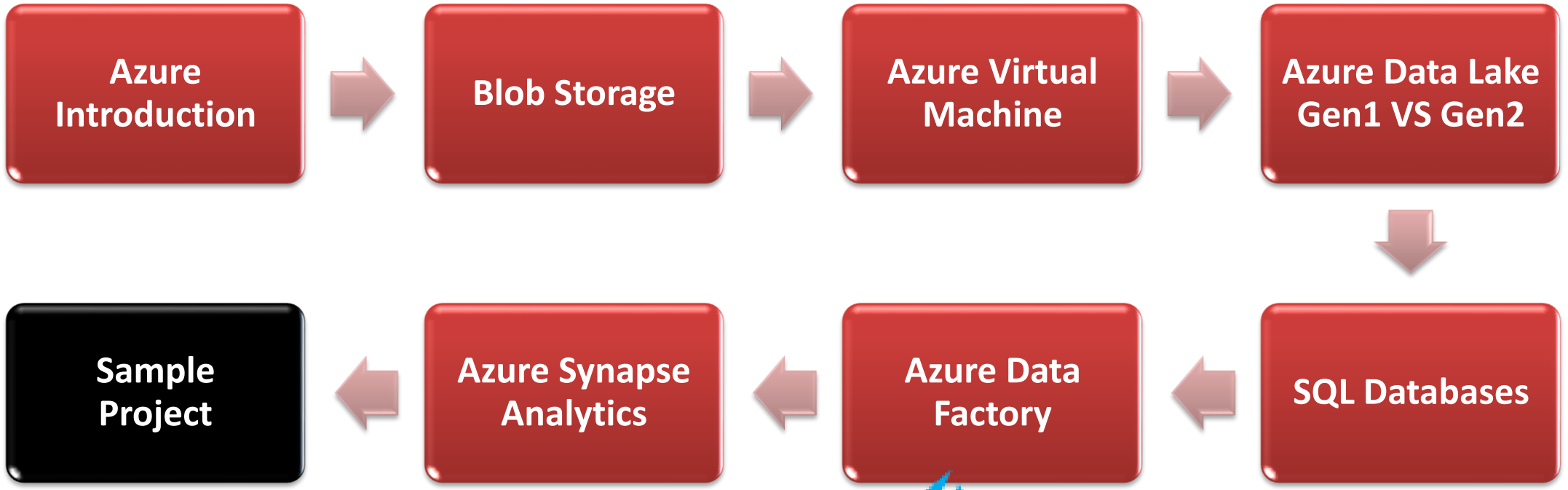
BIGDATAPEDIA
Learning will lead the future...

BigData Ecosystem
Airflow Introduction



BIGDATAPEDIA
Learning will lead the future...

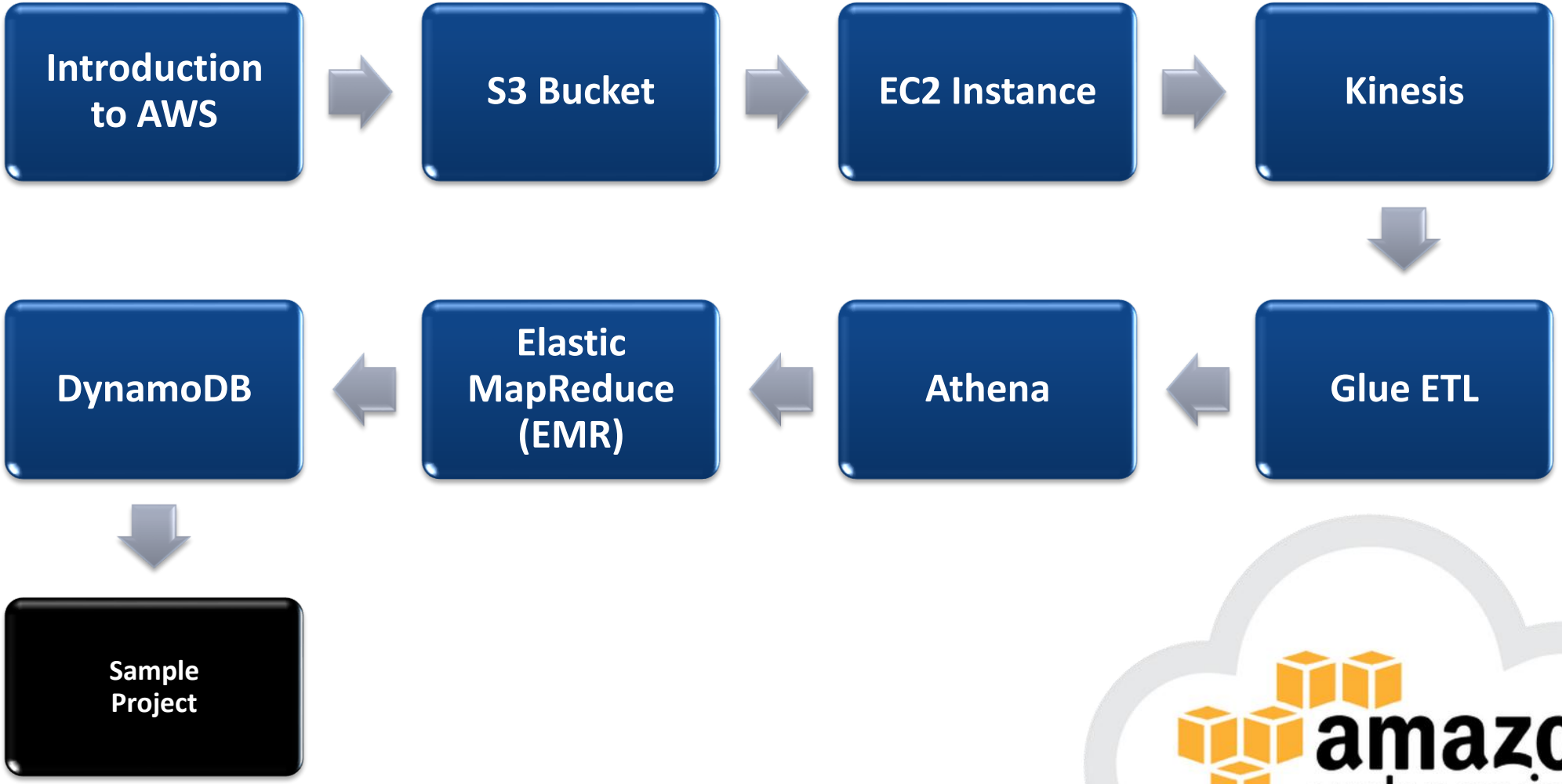
Topic - 2
Azure Services



Azure

BIGDATAPEDIA
Learning will lead the future...

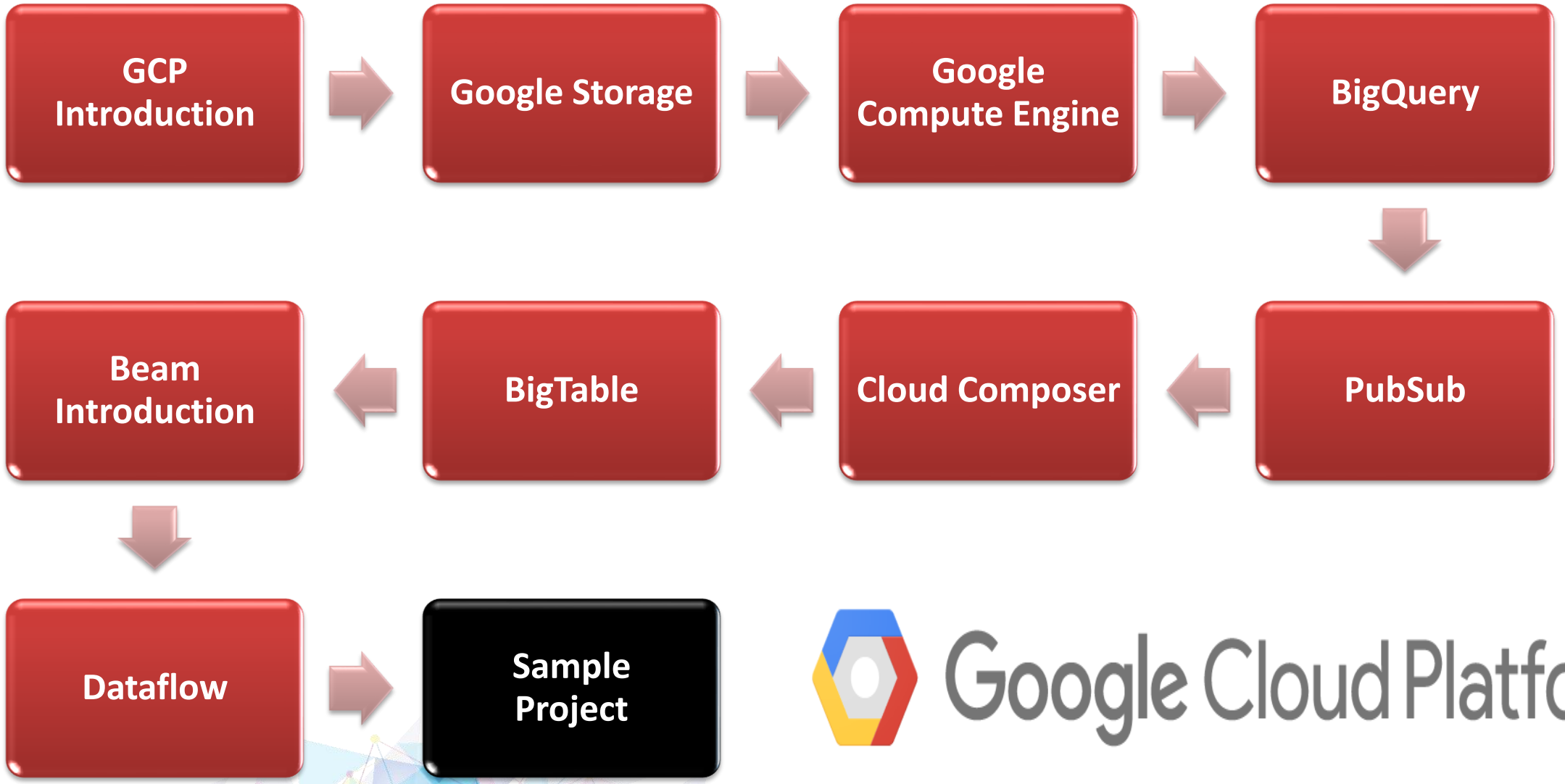
Topic - 3
AWS Services



amazon web services™

BIGDATAPEDIA
Learning will lead the future...

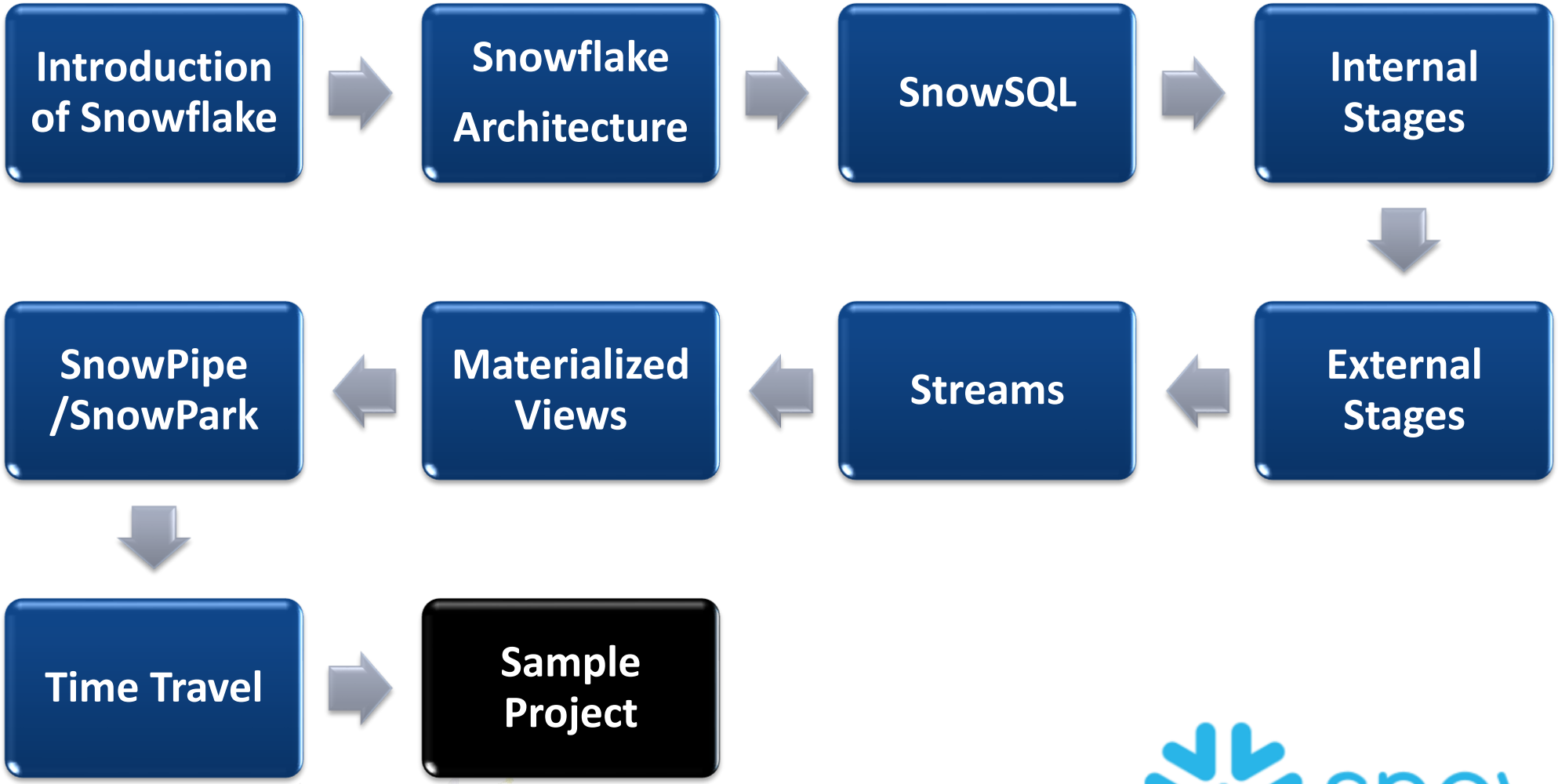
Topic - 4
Google Cloud Services



Google Cloud Platform

BIGDATAPEDIA
Learning will lead the future...

Topic - 5
Snowflake Cloud Data Services

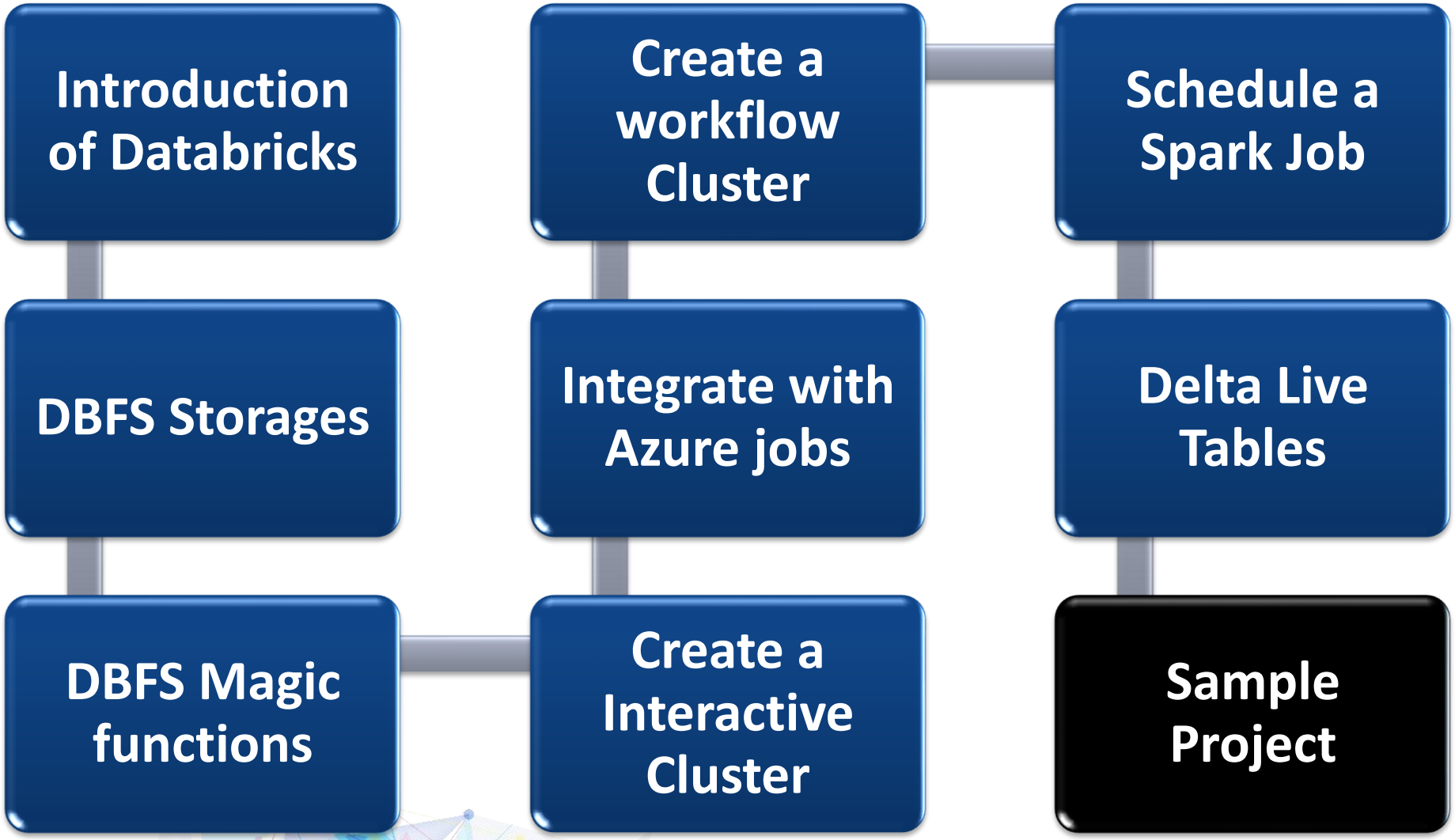


snowflake®

BIGDATAPEDIA
Learning will lead the future...

Topic - 6
Azure Databricks

databricks



- ***Get Free Lifetime Practice with all recent upgraded VM.***
- ***Pay only for the new tools (Only for Old students).***
- ***Bigdata Interview related tips and tricks & Interview question with answers.***

BIGDATAPEDIA

Learning will lead the future...

Our Associate Partners

AXEL ACADEMY

TAMILBOOMI TECHNOLOGIES

www.bigdatapedia.com

Contact us @ +91 9715010010

Please feel free to reach us If you have any queries...

+91 9715 010 010

BigDatapedia ML