CHaracteristics of Big Data

Data Quantity---Volume

Data Speed---Velocity

Data types---Variety

Apache Hadoop software lib is a framework that allows for distributed processing of larger datasets across clusters of computers

Designed to scale up from single servers to thousands providing local computation and storage

For high availability rather than relying on hardware at application level.

Apache Spark Architecture:

The component of data is Resilient Distributed dataset(RDD)

Spark SQL

data sources:avro,paraquet,json,orc

Spark will get installed in databrick cluster

Dataset is converted into dataframe and then to RDD by spark program.

Rdd gets stored in spark core in Databrick

Spark Sql Architecture:

Spark is supported by Api like spark,scala,python

----------Lunch break----------

Features of Spark Sql

1.Integrated

Mix sqlqueries with spark programs

2.Unified Data Access

Load and query data from a variety of sources

3.Hive Compatibility

4.Standard Connectivity

5.Scalability

Same engine for both interactive and long queries

cluster---Group of machines(vm) that runs sparkand its connected to DataWarehouse

Ex: azure databricks,hadoop

setup spark cluster on azure databrick--adb session

spark cluster in opensource

Spark RDD

RDD fundamental data structure of Spark

RDD is read only, partitioned collection of records

It is created through deterministic operations on data or other rdds.

2 methods to create:

1.Parallelizing existing collection in driver program

2.Referencing dataset in external storage system

Usually datasets come from different data sources like storage accounts, lakes, hbase, hdfs (Hadoop distributed file system),fs(file system).These datasets are connected through spark cluster

RDD usage in spark achieves faster and efficient MapReduce operations.

MapReduce algorithm to retrieve faster RDDs

For every word in sentence, It assigns a value

Spark Cluster

Single node spark cluster---one machine with 14 gb ram and 14 gb memory with spark installed

Two node spark Cluster---Two machines with launching cluster with spark installed

got a input of dataset--load the sataset --spark rdd program to convert dataset to RDD

IN ANY node data can be stored.

Data set and DataFrame

A distributed collection of dataorganized into named columns---DataFrame similar to relational tables.

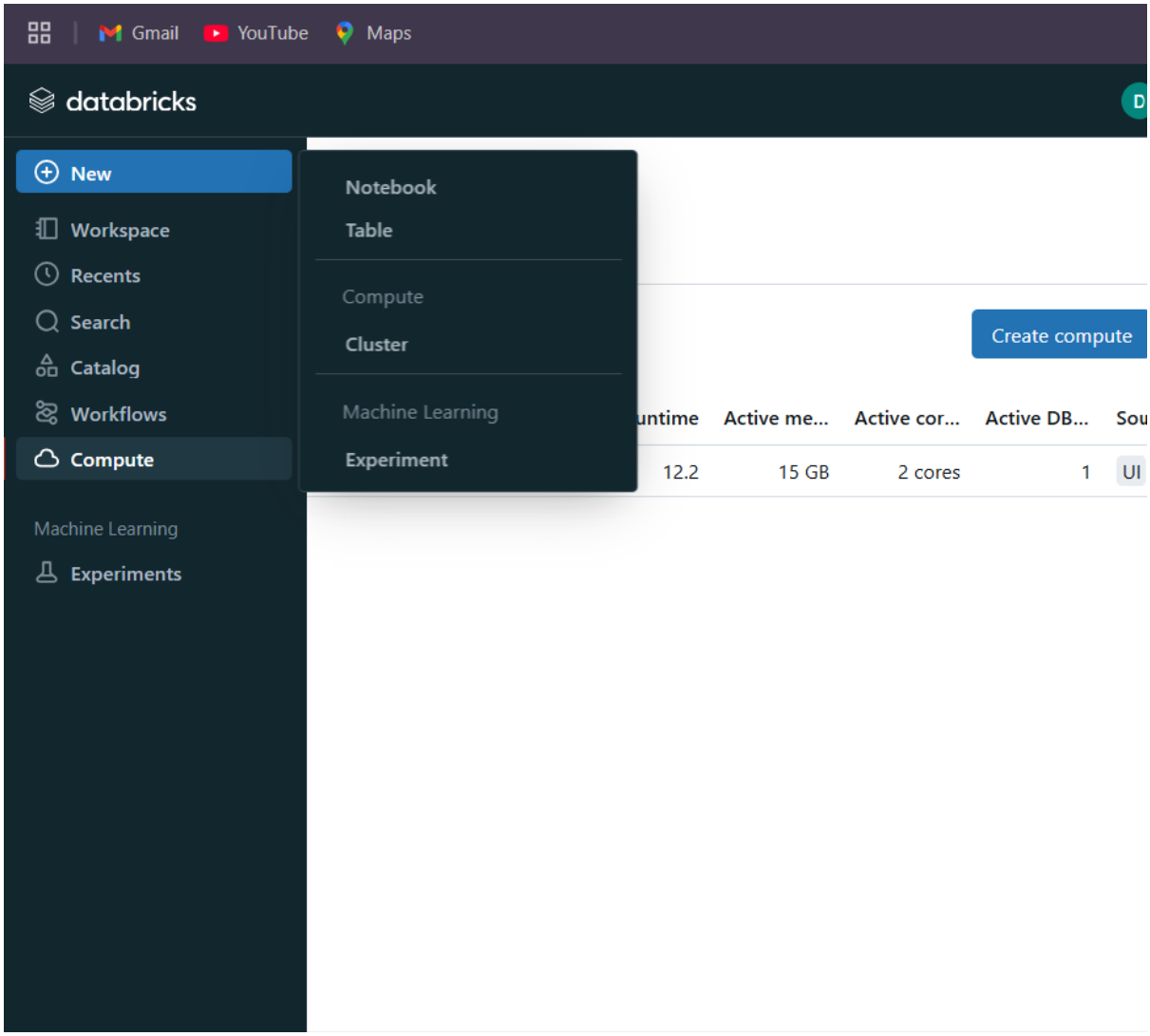
Dataframe can be constructed from array

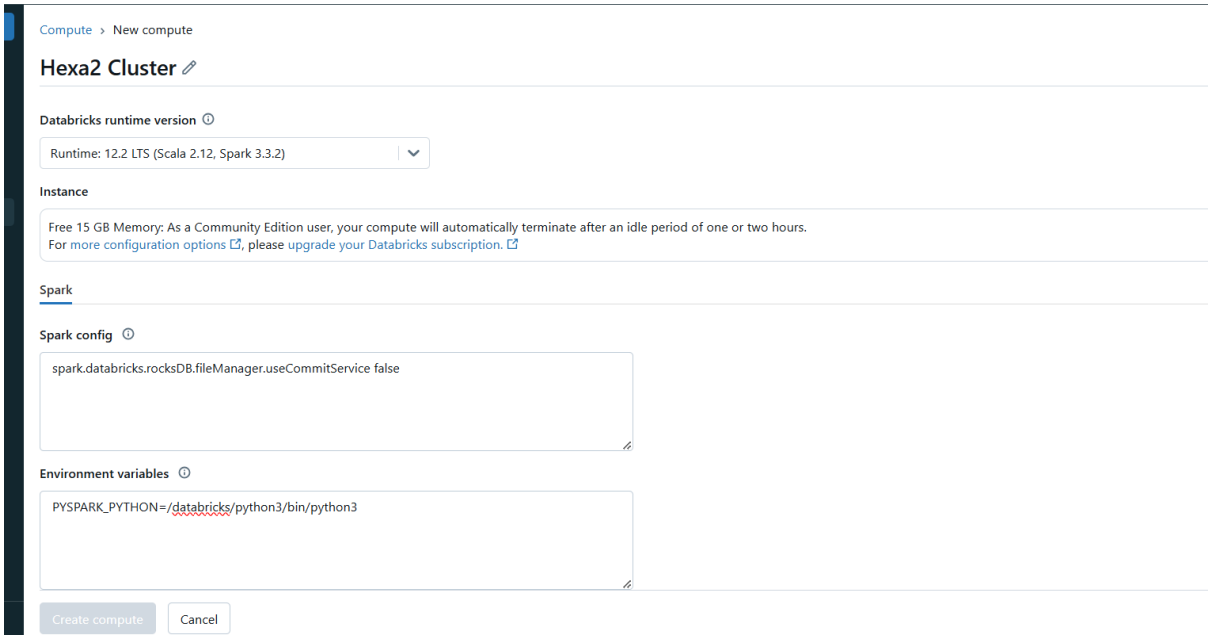
Distributed collection of data--Dataset

dataset--->RDD--->Dataframe

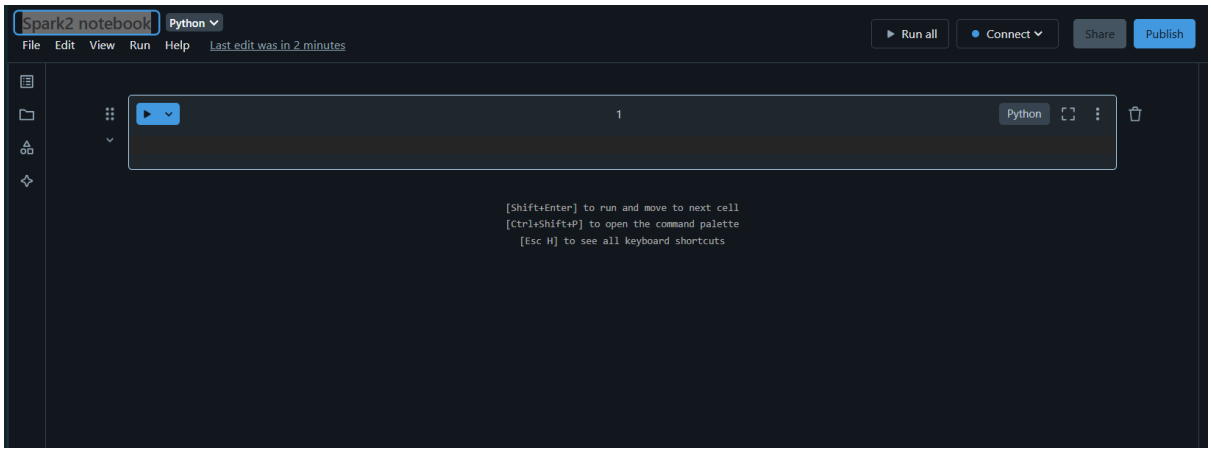
--Features of Dataframe

How to create clusters in databricks

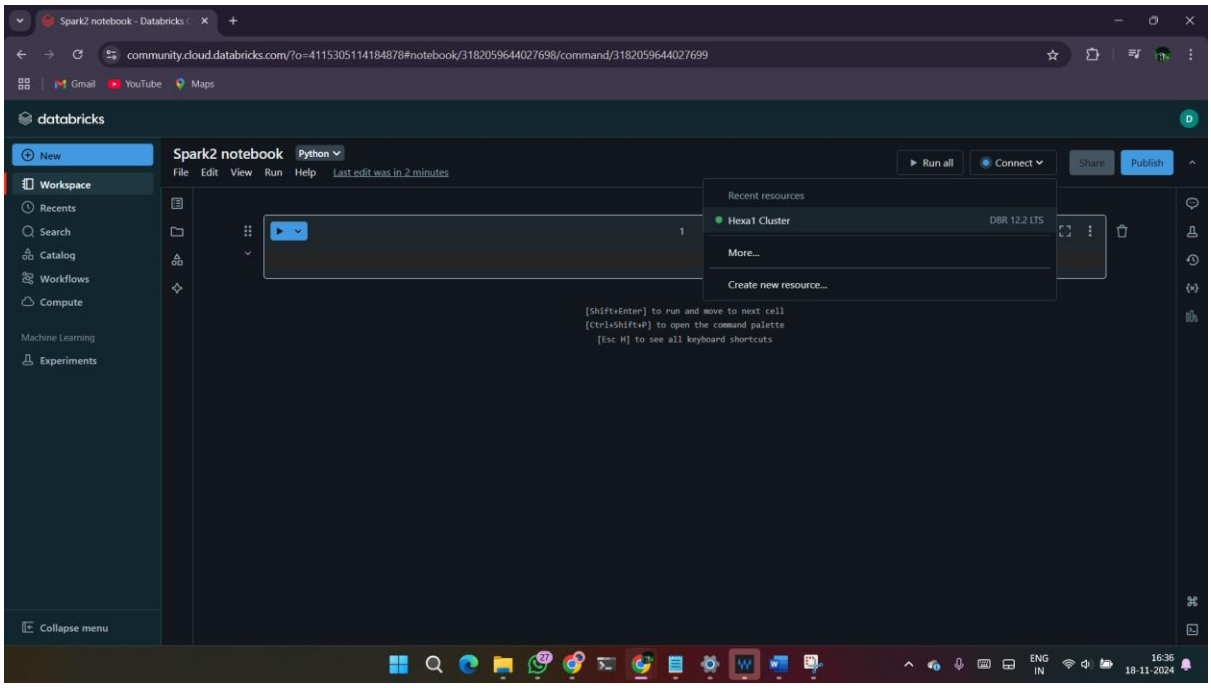

Gmail
YouTube
Maps
databricks
New
Workspace
Recents
Search
Catalog
Workflows
Compute
Machine Learning
Experiments
Notebook
Table
Compute
Cluster
Machine Learning
Experiment
Create compute
untime
Active me...
Active cor...
Active DB...
Sou
12.2
15 GB
2 cores
1
UI


Compute › New compute
Hexa2 Cluster
Databricks runtime version
Runtime: 12.2 LTS (Scala 2.12, Spark 3.3.2)
Instance
Free 15 GB Memory: As a Community Edition user, your compute will automatically terminate after an idle period of one or two hours.
For more configuration options, please upgrade your Databricks subscription.
Spark
Spark config
spark.databricks.rocksDB.fileManager.useCommitService false
Environment variables
PYSPARK_PYTHON=/databricks/python3/bin/python3
Create compute
Cancel

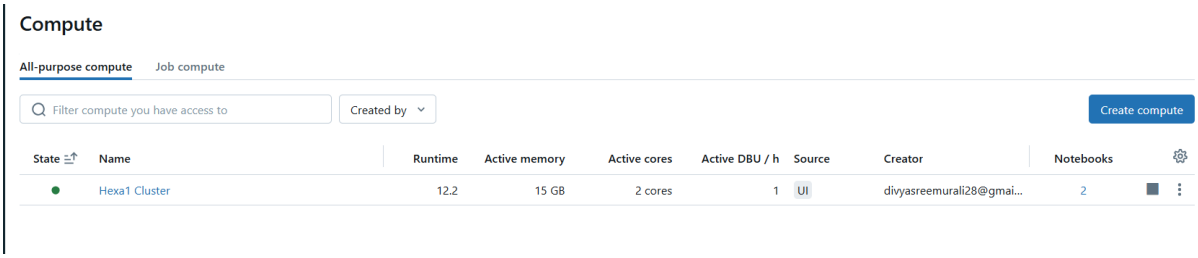
Creating a notebook



Connecting notebook to cluster



In compute, they can be viewed



If required we can delete permanently or terminate permanently.