1. Explain what are various components of SPARK with block diagram?

explain functionality of every components?

Ans- (i)SPARK CORE

Spark Core is, as the name suggests, the core unit of a Spark process. It takes care of task scheduling, fault recovery, memory management, and input-output operations, etc. Think of it as something similar to CPU to a computer. It supports programming languages like Java, Scala, Python, and R and provides APIs for respective languages using which you can build your ETL job or do analytics. All the other components of Spark have their own APIs which are built on top of Spark Core. Because of its parallel processing capabilities and in-memory computation, Spark can handle any kind of workload.

(ii)SPARK SQL

The Spark SQL component is built above the spark core and used to provide the structured processing on the data. It provides standard access to a range of data sources. It includes Hive, JSON, and JDBC. It supports querying data either via SQL or via the hive language. This also works to access structured and semi-structured information. It also provides powerful, interactive, analytical application across both streaming and historical data. Spark SQL could be a new module in the spark that integrates the relative process with the spark with programming API. The main functionality of this module is:

(a)It is a Spark package for working with structured data.

(b)It Supports many sources of data including hive tablets, parquet, json.

(c)It allows the developers to intermix SQK with programmatic data manipulation supported by RDDs in python, scala and java.

(iii)SPARK STREAMING

Spark streaming permits ascendible, high-throughput, fault-tolerant stream process of live knowledge streams. Spark can access data from a source like a flume, TCP socket. It will operate different algorithms in which it receives the data in a file system, database and live dashboard. Spark uses Micro-batching for real-time streaming. Micro-batching is a technique that permits a method or a task to treat a stream as a sequence of little batches of information. Hence spark streaming groups the live data into small batches. It delivers it to the batch system for processing. The functionality of this module is:

(a)Enables processing of live streams of data like log files generated by production web services.

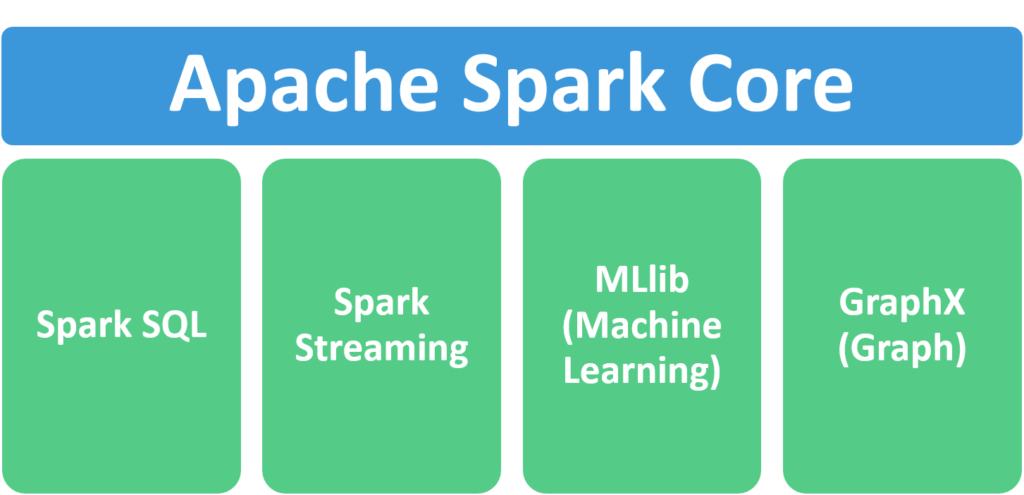
(b)The API’s defined in this module are quite similar to spark core RDD API’s.

(iv) Mlib Machine Learning

MLlib in spark is a scalable Machine learning library that contains various machine learning algorithms. The motive behind MLlib creation is to make the implementation of machine learning simple. It contains machine learning libraries and the implementation of various algorithms. For example, clustering, regression, classification and collaborative filtering.

(v) GraphX graph processing

It is an API for graphs and graph parallel execution. There is network analytics in which we store the data. Clustering, classification, traversal, searching, and pathfinding is also possible in the graph. It generally optimizes how we can represent vertex and edges in a graph. GraphX also optimizes how we can represent vertex and edges when they are primitive data types. To support graph computation, it supports fundamental operations like subgraph, joins vertices, and aggregate messages as well as an optimized variant of the Pregel API.



2. Explain Spark core in details & how RDD is related to Spark core - explain with Spark program ?

Ans- Spark Core is the base of the whole project. It provides distributed task dispatching, scheduling, and basic I/O functionalities. Spark uses a specialized fundamental data structure known as RDD (Resilient Distributed Datasets) that is a logical collection of data partitioned across machines. RDDs can be created in two ways; one is by referencing datasets in external storage systems and second is by applying transformations (e.g. map, filter, reducer, join) on existing RDDs.

The RDD abstraction is exposed through a language-integrated API. This simplifies programming complexity because the way applications manipulate RDDs is similar to manipulating local collections of data.

3. Explain various Mlib algorithms Spark is supporting ?

Ans- The popular algorithms and utilities in Spark MLlib are:

a.Basic Statistics

b.Regression

c.Classification

d.Recommendation System

e.Clustering

f.Dimensionality Reduction

g.Feature Extraction

h.Optimization

a. Basic Statistics

Basic Statistics includes the most basic of machine learning techniques. These include:

Summary Statistics: Examples include mean, variance, count, max, min and numNonZeros.

Correlations: Spearman and Pearson are some ways to find correlation.

Stratified Sampling: These include sampleBykey and sampleByKeyExact.

Hypothesis Testing: Pearson’s chi-squared test is an example of hypothesis testing.

Random Data Generation: RandomRDDs, Normal and Poisson are used to generate random data.

b. Regression

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

c. Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. It is an example of pattern recognition.

Here, an example would be assigning a given email into “spam” or “non-spam” classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

d. Recommendation System

A recommendation system is a subclass of information filtering system that seeks to predict the “rating” or “preference” that a user would give to an item. Recommender systems have become increasingly popular in recent years, and are utilized in a variety of areas including movies, music, news, books, research articles, search queries, social tags, and products in general.

Recommender systems typically produce a list of recommendations in one of two ways – through collaborative and content-based filtering or the personality-based approach.

i. Collaborative Filtering approaches building a model from a user’s past behavior (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in.

ii. Content-Based Filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties.

e. Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). So, it is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics.

f. Dimensionality Reduction

Dimensionality Reduction is the process of reducing the number of random variables under consideration, via obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Feature Selection: Feature selection finds a subset of the original variables (also called features or attributes).

Feature Extraction: This transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in Principal Component Analysis(PCA), but many nonlinear dimensionality reduction techniques also exist.

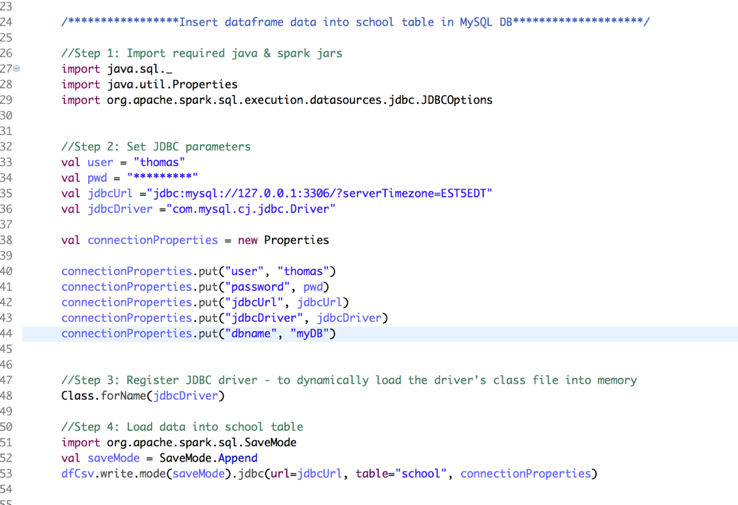
g. Optimization

Optimization is the selection of the best element (with regard to some criterion) from some set of available alternatives.

In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. More generally, optimization includes finding “best available” values of some objective function given a defined domain (or input), including a variety of different types of objective functions and different types of domains.

4. Explain benefits of Spark SQL & how relational data will be inserted into SPARK ?

Ans- It is a module of Apache Spark that analyses the structured data. It provides Scalability, it ensures high compatibility of the system. It has standard connectivity through JDBC or ODBC. Thus, it provides the most natural way to express the Structured Data.



5. Explain Spark streaming in detail ?

Ans- Apache Spark Streaming is an extension of the core Apache Spark API, a distributed general-purpose cluster computing framework that natively supports both batch and streaming workloads. Spark Streaming serves as the entry point for live streaming data, and allows data engineers and data scientists to process real-time data from sources including, but not limited to Kafka, Kinesis, Flume and web APIs such as Twitter. Processed data from these sources can be delivered to file systems, databases, live dashboards or other destinations.

The key abstraction used by Spark Streaming is called a Discretized Stream, or DStream, which in turn is built upon Spark’s key abstraction, RDDs, or resilient distributed datasets. The DStream represents a continuous stream of data, either the input stream from a designated source, or the processed stream generated by transforming the source stream. Each RDD in a DStream contains data from a certain interval. Internally, the flow is as follows:

a.The streamed data is ingested by Spark Streaming.

b.The stream is broken up into small (micro) batches, fed into the Spark engine.

c.The Spark engine processes the data, transforming as necessary

d.Batches of processed data are sent to their designated destinations.

The underlying RDD functionality ensures that Spark Streaming seamlessly integrates with other Spark components such as MLlib (Spark’s Machine Learning Library) and Spark SQL. The key advantage of this is that a single framework can be used to satisfy advanced processing needs.

6. Explain SPARK architecure? what is Master - Slave architecure ?

Ans- Spark Architecture Overview

Apache Spark has a well-defined layered architecture where all the spark components and layers are loosely coupled. This architecture is further integrated with various extensions and libraries. Apache Spark Architecture is based on two main abstractions:

a.Resilient Distributed Dataset (RDD)

b.Directed Acyclic Graph (DAG)

a. The Resilient Distributed Datasets are the group of data items that can be stored in-memory on worker nodes. Here,

Resilient: Restore the data on failure.

Distributed: Data is distributed among different nodes.

Dataset: Group of data.

b. Directed Acyclic Graph is a finite direct graph that performs a sequence of computations on data. Each node is an RDD partition, and the edge is a transformation on top of data. Here, the graph refers the navigation whereas directed and acyclic refers to how it is done.

The Apache Spark framework uses a master-slave architecture that consists of a driver, which runs as a master node, and many executors that run across as worker nodes in the cluster. Apache Spark can be used for batch processing and real-time processing as well.

7. Explain various cluster managers in SPARK?

a. Standalone Cluster Manager

It is a part of spark distribution and available as a simple cluster manager to us. Standalone cluster manager is resilient in nature, it can handle work failures. It has capabilities to manage resources according to the requirement of applications.

We can easily run it on Linux, Windows, or Mac. It can also access HDFS (Hadoop Distributed File System) data. This is the easiest way to run Apache spark on this cluster. It also has high availability for a master.

b. Hadoop Yarn

This cluster manager works as a distributed computing framework. It also maintains job scheduling as well as resource management. In this cluster, masters and slaves are highly available for us. We are also available with executors and pluggable scheduler.

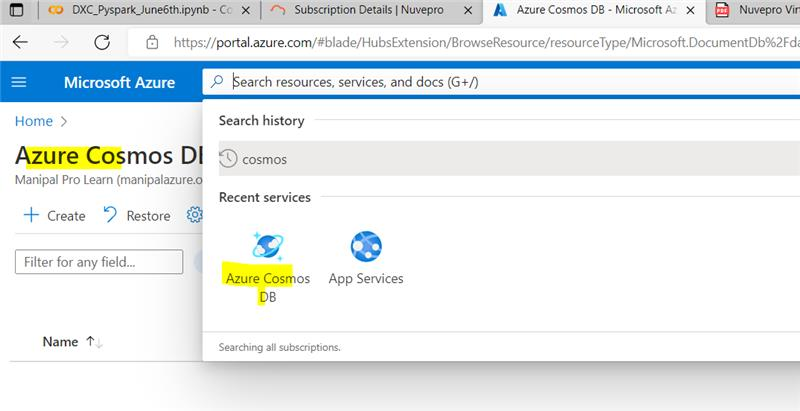
We can also run it on Linux and even on windows. Hadoop yarn is also known as MapReduce 2.0. It also bifurcates the functionality of resource manager as well as job scheduling.

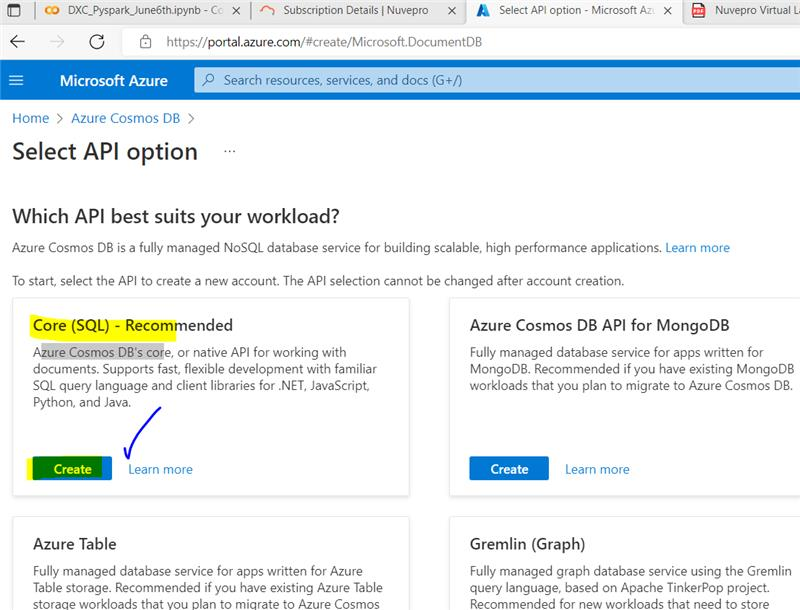
c. Apache Mesos

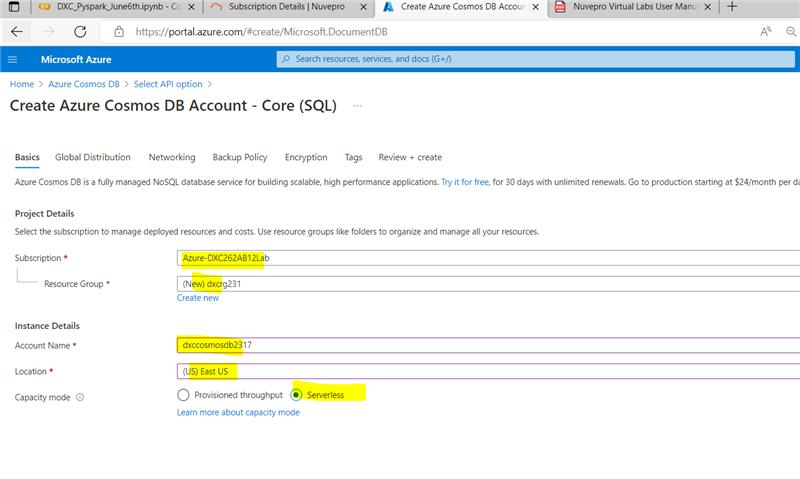
It is a distributed cluster manager. As like yarn, it is also highly available for master and slaves. It can also manage resource per application. We can run spark jobs, Hadoop MapReduce or any other service applications easily.

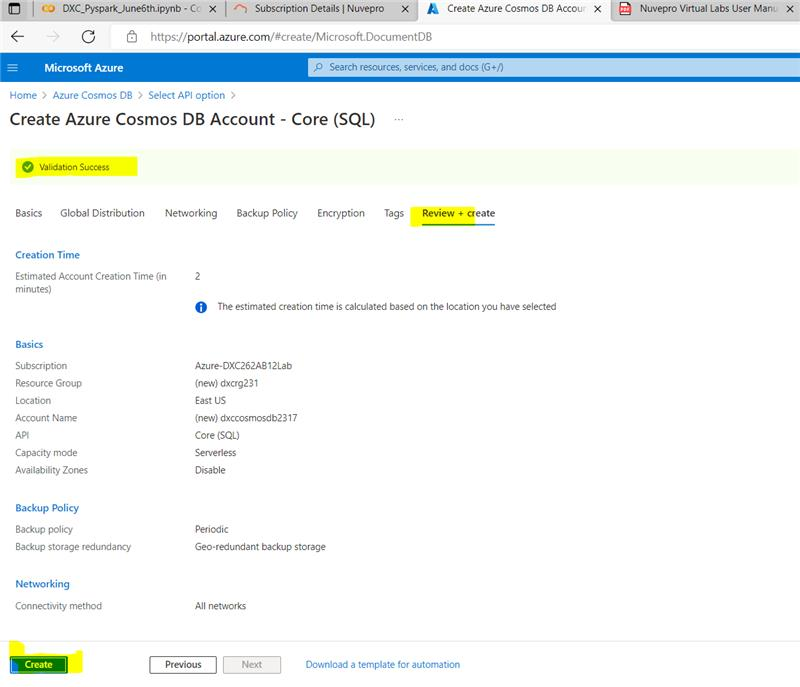
Apache has API’s for Java, Python as well as c++. We can run Mesos on Linux or Mac OSX also.

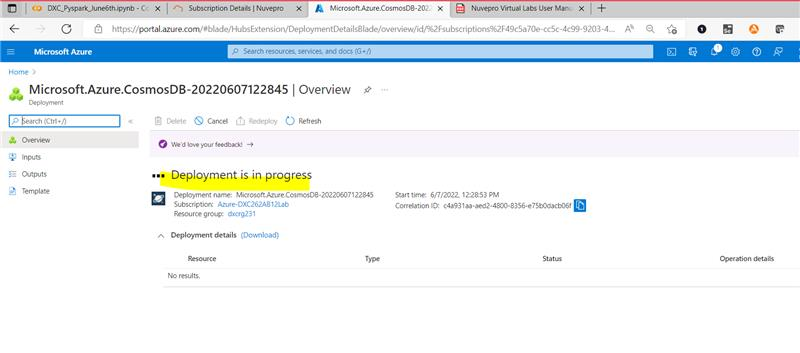


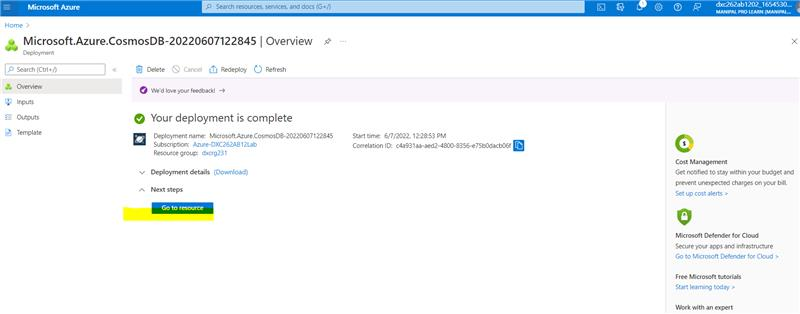




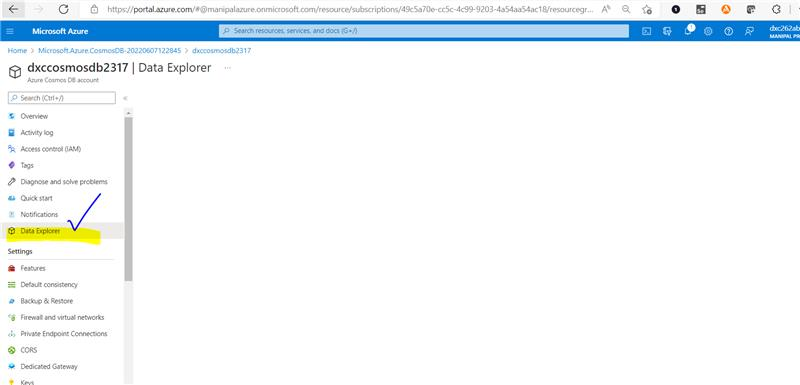


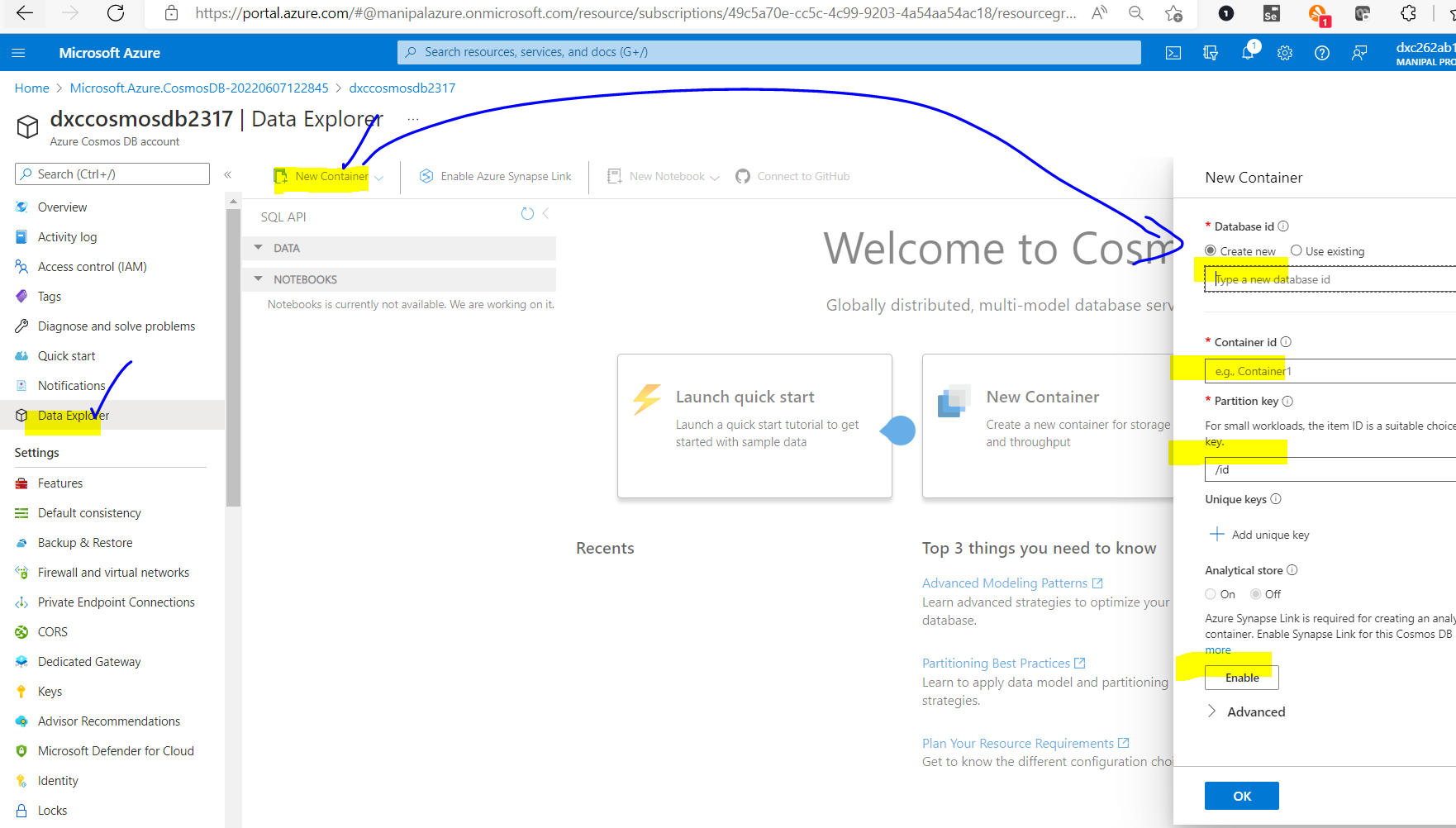


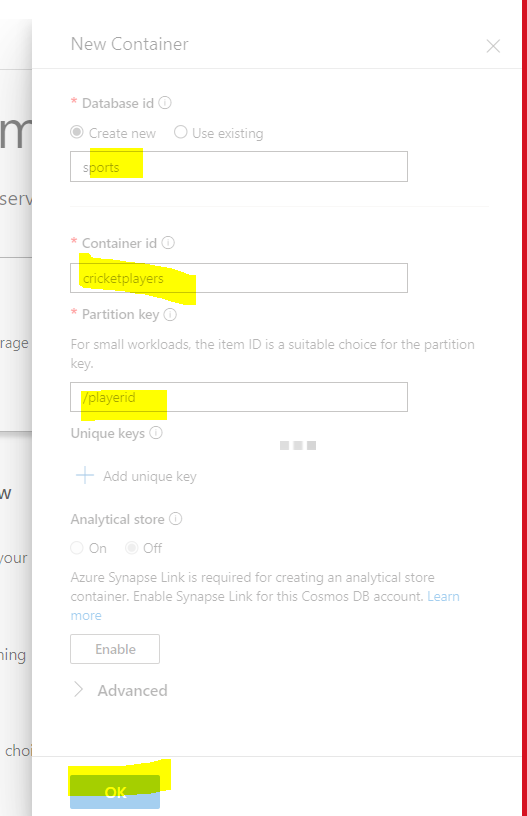


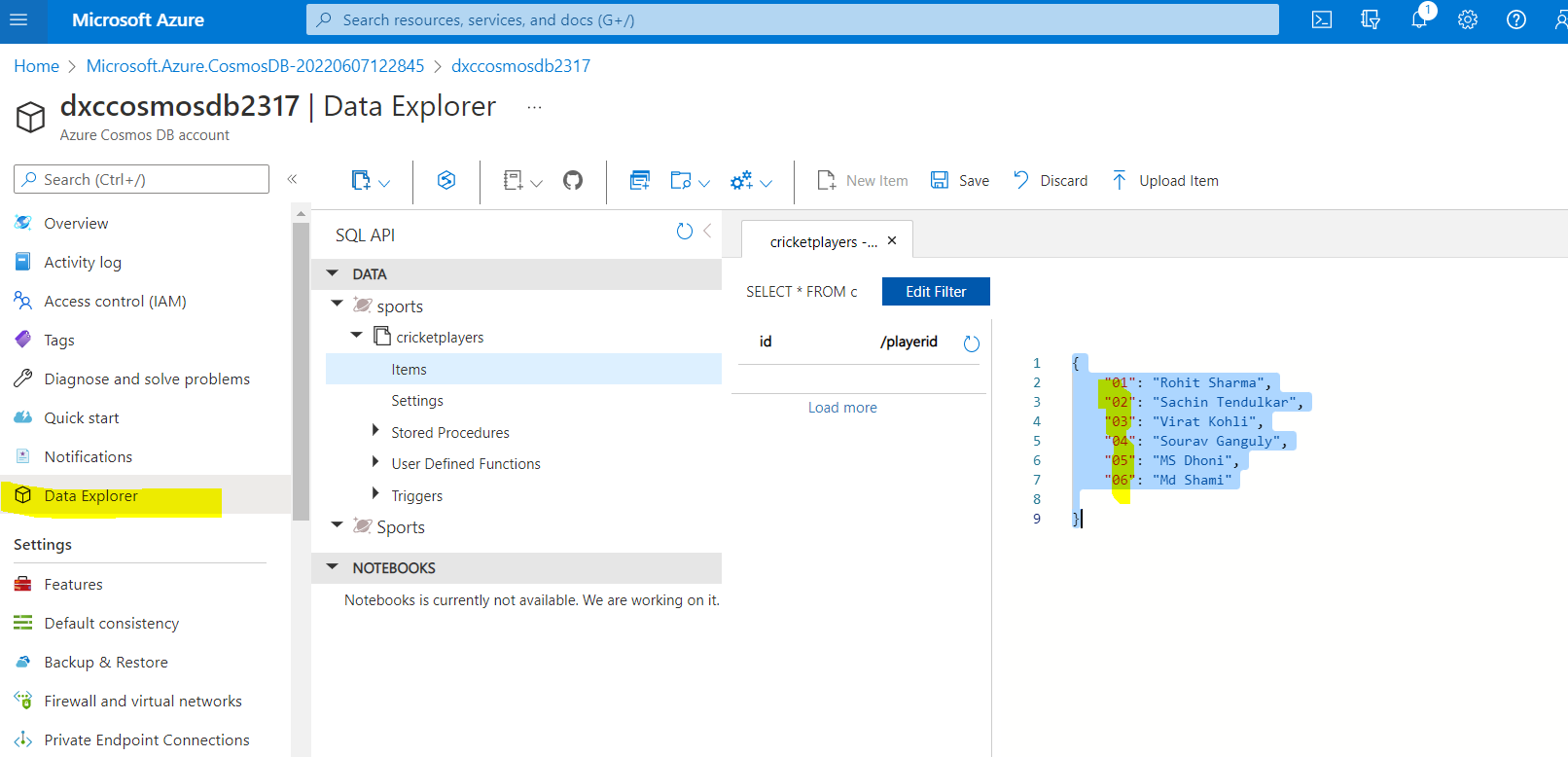


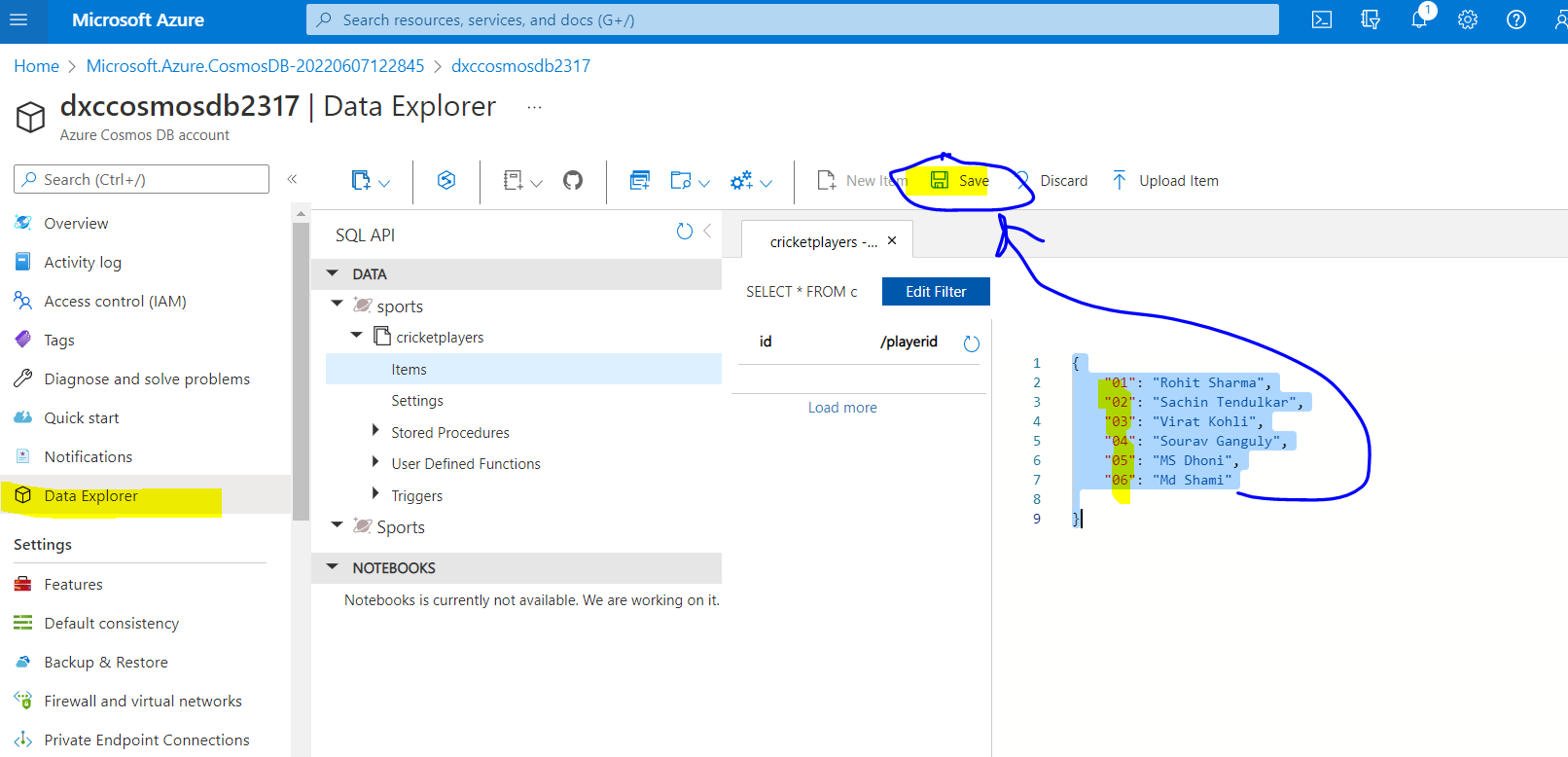
9.

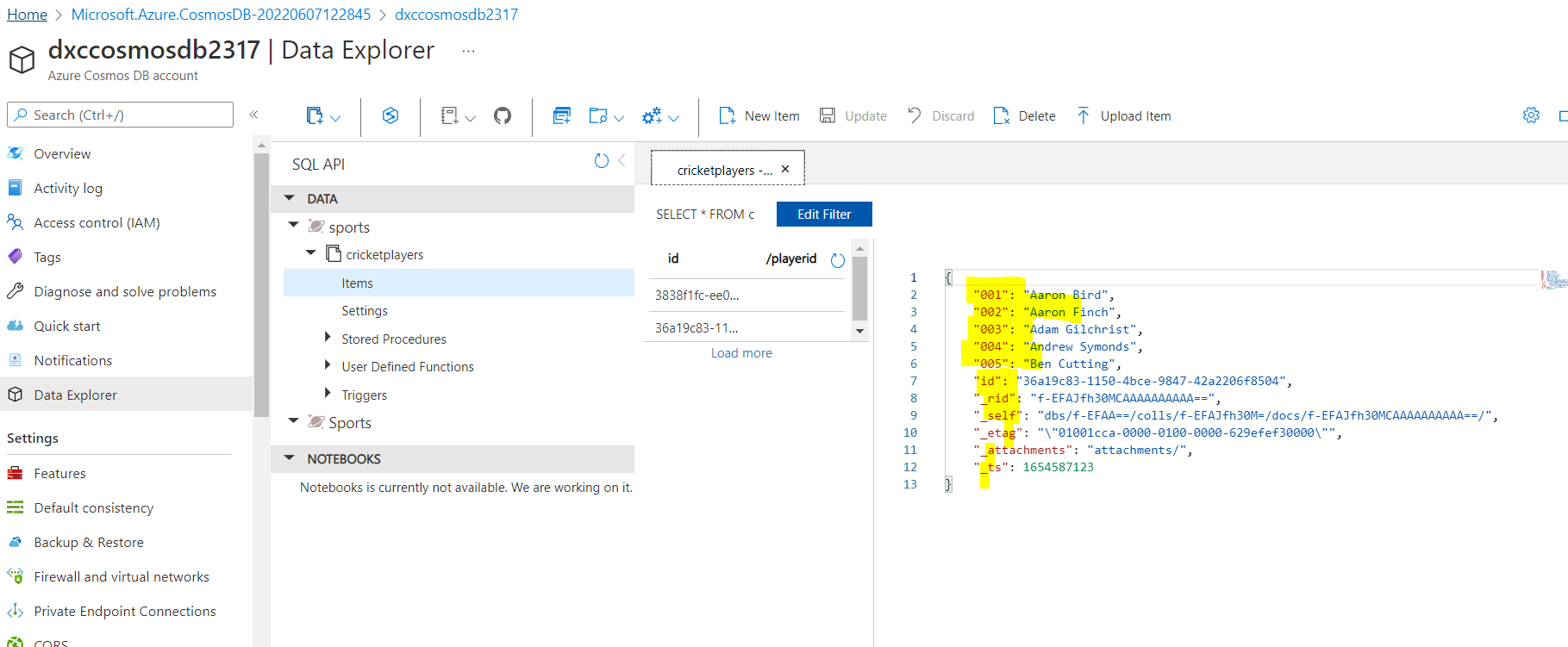












10. Explain with sceenshots & step how to create Azure SQL Db & also explain how to

insert data into Azure SQL D?

