18DCS007
RUDRA BARAD

DSA
UNIT TEST - I
03/09/2021

Q1. Formula for covariance $\text{Cov}(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Mean of x = (98 + 87 + 90 + 85 + 95 + 75) 6 = 88.33

Mean of y = (15 + 12 + 10 + 10 + 16 + 7) /6 = 11.67

Subtract each value from its respective mean & then multiply these new values together

TEMPERATURE $(x - \bar{x})$	CUSTOMER $(y - \bar{y})$	PRODUCT $(x - \bar{x})(y - \bar{y})$
9.67	3.33	32.20
-1.33	0.33	-0.44
1.67	-1.67	-2.79
-3.33	-1.67	5.56
6.67	4.33	28.88
-13.33	-4.67	62.25

Adding all the products together it yields the value 125.66

Final step, divide by (n-1) = 6-1 = 5

∴ 125.66 / 5 = 25.132

Q2. HADOOP is a software framework from apache software foundation that is used to store & process big data.

It is like a platform or a suite which provides various services to solve the big bata problems.

→ In includes open source projects as well as a complete range of complementary tools.

→ Major component of Hadoop include HDFS, YARN, HADOOP COMMON, HIVE, etc-

HDFS - HDFS stands for Hadoop distributed file system. It is major component of Hadoop ecoystm & is responsible for storing large data sets of structured or unstructred data access various node & there by maintaing the metadata in form of log files

WORKING OF HDFS

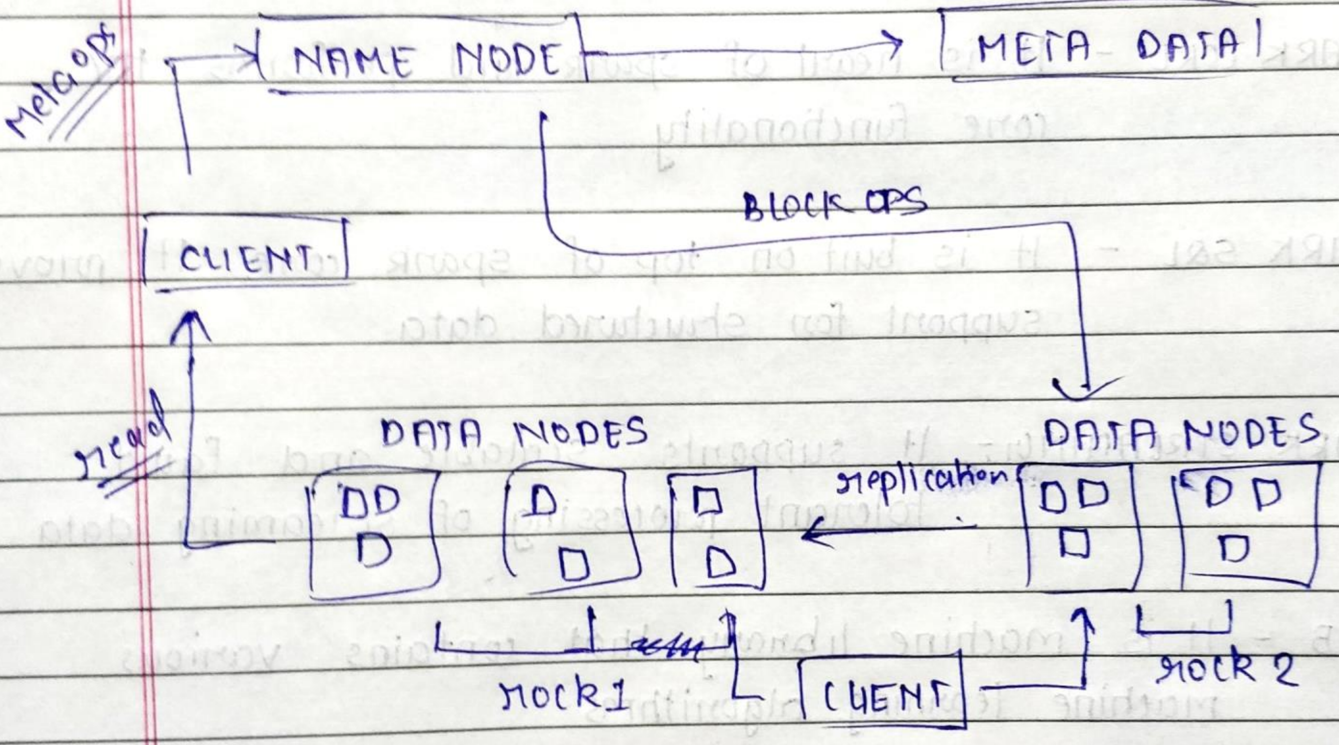
→ It enables rapid transfer of data between nodes.

→ It is closely coupled with map reduce & it organizes & condenses the result into a cohesive answers to query

→ It breaks the info into separate blocks & distributed them to different nodes in cluster.

→ It keeps the track of where file data is kept in the cluster.

HDFS ARCHITECTURE



18DCS007
RUDRA BARAD

Q3 SPARK ECOSYSTEM consists of different types of tightly integrated components. At its core, spark is a computational engine that can schedule, distritute and monitor multiple applications.

SPARK CORE - It is heart of spark and performs the core functionality

SPARK SQL - It is buit on top of spark core. It provide support for structured data

SPARK STREAMING - It supports scalable and fault-tolerant processing of streaming data

MLIB - It is machine library that contains various machine learning algorithms

GRAPHX - It is library that is used to manipulate graphs & perform graph parallel computations.

Features of spark are:

→ Lightning fast processing speed
→ ease of use
→ real time stream processing
→ flexible
→ offeres support for sophisticated analysis
→ active & expanding community