Anay Shah
60004210073
TE Comps C12

5

BDI Assignment 2

Q] from pyspark.sql import SparkSession
spark = SparkSession.builder\.appName("Spark")\.getOrCreate()
url = "Breast_cancer_data.csv"
df = spark.read.csv(url, header=False, inferSchema=True)
columns = ["mean_radius", "mean_texture", "mean_perimeter", "mean_area",
"mean_smoothness", "target"]
df = df.toDF(*columns)
df.createOrReplaceTempView("cancer")

result1 = spark.sql("SELECT * FROM cancer")
result2 = spark.sql("SELECT * FROM cancer WHERE mean_radius > 15")
result3 = spark.sql("SELECT * FROM cancer WHERE target = 1")
result4 = spark.sql("SELECT COUNT(*) FROM cancer WHERE mean_texture > 15
and mean_perimeter > 100")
result5 = spark.sql("SELECT AVG(mean_area) FROM cancer WHERE target = 0")

result1.show()
result2.show()
result3.show()
result4.show()
result5.show()

spark.stop()

Q2] 5 industry use cases of MongoDB are:-

i) Finance and Banking - MongoDB is used in finance and banking for its ability to handle large volumes of data and support complex transactions. It's employed for real-time fraud detection risk management, compliance reporting and personalized customer experiences.

ii) Telecommunications :- Telecommunications companies utilizes MongoDB ~~to~~ as it enable telecom providers to analyze customer behaviour, optimize network performance and deliver personalized services. It's also used for managing customer billing, customer relationship management (CRM), and network analytics.

iii) Healthcare - In the healthcare industry, MongoDB is used for managing electronic health records, medical imaging data, patient information and clinical trial data. MongoDB's ability to handle unstructured and semi-structured data makes it suitable for storing diverse healthcare data types.

iv) E-commerce and Retail - MongoDB powers e-commerce platform and retail applications for managing product catalogs, inventory management, order processing and customer interactions. MongoDB's flexible schema allows retailers to adapt to changing product attributes, pricing models & customer preferences.

v) Technology and Software - MongoDB is widely used in the technology and software industry for building modern, data-intensive applications. MongoDB's abilities and features make it the perfect fit for several use cases and scenarios.

Q3) 5 industry use cases for Apache Kafka:-

i) Internet of Things (IoT) - Apache Kafka serves as a central component in IoT platforms for ingestion, processing and analyzing streaming data from connected devices such as sensors, actuators, wearables and industrial machines.

ii) Media and entertainment - Kafka enables media organisations to build personalized content delivery platforms, real-time analytics dashboards and audience engagement tools.

iii) ~~Transporati~~ Transportation and Logistics - In the transportation and logistic industry, Apache Kafka is used for managing real-time data from GPS devices, telematics systems, sensors and supply chain tracking systems.

iv) Energy and Utilities - Energy companies utilize Apache Kafka for processing streaming data from smart meters, IoT devices, energy grids and renewable energy sources. Kafka enables real-time ~~non~~ monitoring of energy consumption, grid stability and power generation.

v) Gaming - Apache Kafka is used for real-time processing of game events, player interactions, in-game purchases, and multiplayer game sessions. Kafka enables game developers to build scalable multiplayer gaming platforms, real-time leaderboards, match-making systems and social gaming features.