Intro:

======

Hello, my name is sheshank. I've been working from more than eight plus years in IT Industry and 5 + years as a data engineer. I worked on designing, developing, ingesting data from multiple data sources and transform the data and put that into the data format so that downstream users can create dashboards and reports out of it. Mostly working with JSON files, CSV files, XML's and then converted into parquet format. I've been involved in designing developing pipelines, end to end batch and streaming. For data streaming, we have been using Kafka, and then build couple of producers and consumers. Also worked on migrating the pipelines from on-prem to cloud. on-prem I have worked with Hadoop, Hive spark.

And also on a daily basis, I work on Python, SQL and Kafka and worked on building and them deploying. I've been involved in developing AWS specific solutions using s3 redshift, glue, EMR, SQS, SNS, lambda and Terraform script.

let me explain about my current project. We are receiving the data in s3 and then from the source system, we basically need to convert that into parquet format. So, we instantiate an EMR cluster and then in EMR, we have different steps. In step three, create transformation logic. We basically build our transformations by reading JSON files using Spark and then we convert that into parquet.

Utilized Apache Spark on Data bricks to perform data processing and analysis on large-scale datasets, and optimized Spark jobs for performance and scalability.

Worked with AWS services such as S3, Redshift, and Glue to store and manage data, and integrated them with Data bricks pipelines for data processing and transformation.

Designed and implemented data validation and quality checks using Data bricks, ensuring that data was accurate and consistent across multiple sources.

And then in the previous project, we also worked on GCP and services like GCS, cloud composer, data flow and big query. So, using the query created different views and stored procedures and records out of it. And then from composite standpoint, we received the data from different stakeholders in CSVs the properties and them and then the using different operators, we could insert job operator. We are Kubernetes pod operator; we process our data using airflow composer. and then we store that data into the Big Query. So, I've been involved in setting up buckets IAM policies, roles and then creating Big Query data.

* Develop Shell script that reads Json files and apply it to Sqoop and Hive.
* Ingested data from Relational DB, Oracle database, Postgres SQL using SQOOP into HDFS and loaded them into Hive tables, AWS S3, transformed and analyzed large datasets by running Hive queries and using Apache Spark.
* Work with pyspark to migrate fixed width, ORC, csv etc. files.
* Designed and implemented an ETL framework to load data from multiple sources into Hive and from Hive into Teradata.
* Utilized SQOOP, ETL and Hadoop File System API's for implementing data ingestion pipelines.
* Worked on Batch data of different granularity ranging from hourly, daily to weekly and monthly.
* Handled Hadoop cluster installations in various environments such as Unix, Linux, and Windows
* Assisted in upgrading, configuration, and maintenance of various Hadoop infrastructures like Ambari, spark and Hive.
* Work with stream sets and develop pipelines using stream sets
* Developing and writing SQLs and stored procedures in Teradata. Loading data into snowflake and writing Snow SQLs scripts
* TDCH scripts for full and incremental refresh of Hadoop tables.
* Optimizing Hive queries by parallelizing with portioning and bucketing.
* Worked on various data formats like AVRO, Sequence File, JSON, Map File, Parquet and. ORC.
* Worked extensively on Teradata, Hadoop-Hive, Spark, SQLs, PLSQLs
* Designed and published visually rich and intuitive Stream sets pipelines to migrate data
* Experienced in working with SQL, T-SQL, PL/SQL scripts, views, indexes, stored procedures, and other
* components of database applications
* Experienced in working with Hadoop from Horton works Data Platform and running services through Cloudera manager
* Used Agile Scrum methodology/ Scrum Alliance for development

Environment: Hadoop, HDFS, AWS, Vertica, Scala, Kafka, MapReduce, YARN, Spark, Hive, Scala, MySQL, Kerberos, Maven, Stream sets.

How much data you handled:

200 GB of data on daily basis.

Cluster size:

on prem 250+ nodes cluster team size of

Sources of your data:

RDBMS systems like MySQL, Teradata

Kafka stream

file formats you handled:

json, csv, xml, Avro, and parquet.

Services in AWS:

S3, Glue, EMR, redshift, sns, sqs, lambda

Services in GCP:

GCS, data proc, dataflow, cloud composer, big query

services in azure:

Azure data lake storage (ADLS), ADF (azure data factory), Azure data bricks

1. Database Design: Designing and creating databases, tables, views, indexes, and relationships using SQL.
2. Data Integration: Importing data from various sources into the database using SQL scripts.
3. Data Transformation: Writing SQL scripts to transform data, such as converting data types, splitting columns, and aggregating data.
4. Data Quality: Implementing data validation and data cleaning tasks using SQL.
5. Data Processing: Automating data processing tasks such as data extraction, transformation, and loading, using SQL scripts.
6. Reporting and Analysis: Creating reports and dashboards using SQL to retrieve and aggregate data from the database.
7. Performance Optimization: Optimizing SQL queries and database performance by indexing, partitioning, and tuning SQL scripts.
8. Technical Support: Providing technical support to business users and other stakeholders, including resolving technical issues, answering questions, and providing training on the use of SQL.
9. Database Maintenance: Performing database maintenance tasks, such as backing up databases, monitoring database performance, and resolving database issues.
10. Collaboration: Collaborating with other developers, data analysts, and stakeholders to understand requirements, develop solutions, and ensure the quality of the data.
11. Writing TSQL scripts to transform data, such as converting data types, splitting columns, and aggregating data.
12. Data Quality: Implementing data validation and data cleaning tasks using TSQL.
13. Data Processing: Automating data processing tasks such as data extraction, transformation, and loading, using TSQL scripts.
14. Reporting and Analysis: Creating reports and dashboards using TSQL to retrieve and aggregate data from the SQL Server database.

Data modelling is **the process used to structure how data is stored, as well as modelling relationships within the data**. The goal is to create a visual data map that accurately describes the data structure, how data will flow through the system whilst highlighting important data relationships.

Azure Synapse Analytics is a limitless analytics service that brings together data integration, enterprise data warehousing, and big data analytics. It gives you the freedom to query data on your terms, using either serverless or dedicated resources—at scale. Azure Synapse brings these worlds together with a unified experience to ingest, explore, prepare, transform, manage, and serve data for immediate BI and machine learning needs.

**PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing**. If you're already familiar with Python and libraries such as Pandas, then PySpark is a good language to learn to create more scalable analyses and pipelines.

There are many ETL tools available in the market that can carry out this process. **A standard ETL tool like PySpark**, supports all basic data transformation features like sorting, mapping, joins, operations, etc. PySpark's ability to rapidly process massive amounts of data is a key advantage.

Develop deep understanding of the data sources, implement data standards, maintain data quality and master data management.  
Expert in developing JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data.  
Expert in using Databricks with Azure Data Factory (ADF) to compute large volumes of data.  
Performed ETL operations in Azure Databricks by connecting to different relational database source systems using jdbc connectors.  
Developed Python scripts to do file validations in Databricks and automated the process using ADF.  
Developed an automated process in Azure cloud which can ingest data daily from web service and load into Azure SQL DB.  
Developed Streaming pipelines using Azure Event Hubs and Stream Analytics to analyze data for dealer efficiency and open table counts for data coming in from IOT enabled poker and other pit tables.  
Analyzed data where it lives by Mounting Azure Data Lake and Blob to Databricks.  
Used Logic App to take decisional actions based on the workflow.  
Developed custom alerts using Azure Data Factory, SQLDB and Logic App.  
Developed Databricks ETL pipelines using notebooks, Spark Data frames, SPARK SQL and python scripting.

O Design and implement database solutions in Azure SQL Data Warehouse, Azure SQL.

• Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NOSQL DB).

O Design & implement migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party

tools.

O Engage with business users to gather requirements, design visualizations and provide training to use self-service BI tools.

• Used various sources to pull data into Power BI such as SQL Server, Excel, Oracle, SQL Azure etc.

• Propose architectures considering cost/spend in Azure and develop recommendations to right-size data infrastructure.

O Develop conceptual solutions & create proof-of-concepts to demonstrate viability of solutions.

O Technically guide projects through to completion within target timeframes.

O Collaborate with application architects and DevOps.

• Identify and implement best practices, tools and standards.

• Design Setup maintain Administrator the Azure SQL Database, Azure Analysis Service, Azure SQL Data warehouse, Azure Data Factory, Azure SQL Data warehouse.

O Build Complex distributed systems involving huge amount data handling, collecting metrics building data pipeline, and Analytics.

As someone with a working knowledge of Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) models, you should have an understanding of the key differences and similarities between these two models.

PaaS:

1. A platform that provides an environment for developing, running, and managing applications.
2. Offers a complete infrastructure and development tools, which reduces the amount of time and effort required to develop and maintain applications.
3. Developers use the platform to build, test, and deploy their applications, while the provider of the platform manages the underlying infrastructure and provides ongoing maintenance and support.

SaaS:

1. A software delivery model in which a software application is hosted by a third-party provider and made available to customers over the internet.
2. Customers do not need to install or maintain the software on their own computers or servers. Instead, they can access the application via a web browser or an API.
3. The provider of the SaaS application is responsible for managing and maintaining the software, including updates, security, and backup and recovery.

Both PaaS and SaaS are designed to reduce the amount of time, effort, and cost involved in developing and deploying software applications, making it easier for organizations to focus on their core business operations. Understanding these models is crucial for businesses, as it can help them make informed decisions about the best way to deploy and manage their applications, depending on their specific needs and requirements.

As a data engineer working with Google Cloud services, your responsibilities may include:

1. Designing and implementing data processing and storage solutions using Google Cloud technologies such as Google Cloud Storage, Google BigQuery, and Google Cloud Dataflow.
2. Migrating data from on-premises systems to Google Cloud, using tools such as Google Cloud Storage Transfer Service and Google BigQuery Data Transfer Service.
3. Designing and implementing data pipelines to automate the movement of data from various sources into Google Cloud.
4. Implementing data access control and security measures to ensure that data is protected and secure.
5. Developing and implementing monitoring and alerting solutions to ensure the health and performance of data processing and storage systems.
6. Collaborating with data scientists and other stakeholders to understand their data requirements and to provide solutions that meet those requirements.
7. Optimizing the performance and cost-effectiveness of data processing and storage solutions.
8. Staying up-to-date with the latest Google Cloud technologies and trends to continuously improve the data engineering processes and solutions.
9. Troubleshooting and debugging any issues that arise during the implementation and operation of data processing and storage solutions.

Google Cloud provides a comprehensive set of data processing and storage services that can help organizations to store, process, and analyze large amounts of data in a scalable and cost-effective manner. As a data engineer, you would play a key role in designing and implementing these solutions, ensuring that the data being processed and managed is secure, accurate, and of high quality.

As an Azure data engineer, your responsibilities may include:

1. Designing and implementing data processing and storage solutions using Azure technologies such as Azure Data Lake Storage, Azure Databricks, and Azure Stream Analytics.
2. Migrating data from on-premises systems to Azure, using tools such as Azure Data Factory, Azure Database Migration Service, and Azure Migrate.
3. Designing and implementing data pipelines to automate the movement of data from various sources into Azure.
4. Implementing data access control and security measures to ensure that data is protected and secure.
5. Developing and implementing monitoring and alerting solutions to ensure the health and performance of data processing and storage systems.
6. Collaborating with data scientists and other stakeholders to understand their data requirements and to provide solutions that meet those requirements.
7. Optimizing the performance and cost-effectiveness of data processing and storage solutions.
8. Staying up-to-date with the latest Azure technologies and trends to continuously improve the data engineering processes and solutions.
9. Troubleshooting and debugging any issues that arise during the implementation and operation of data processing and storage solutions.

Azure provides a comprehensive set of data processing and storage services that can help organizations to store, process, and analyze large amounts of data in a scalable and cost-effective manner. As an Azure data engineer, you would play a key role in designing and implementing these solutions, ensuring that the data being processed and managed is secure, accurate, and of high quality.