Data Bricks

------

Data Bricks

<https://youtu.be/WmiSzxvkA74?si=rigNsgkZ_iY5eWMP> - Databricks

<https://youtu.be/Knpn8yNMjeA?si=mWFfjJK-Y_kEDC4R> – Azure data factory interview questions

ACID transactiosn

Atomicity - if process failed while updating the data. It will allow only complete row is updated or Keep the complete row in the original format

Consistency - Data is observed simultaneous process

Isolation – Querying the table same time

Durability – Committed the changes are permanent

In the main panel we can see the below optiosn

Data science & Engineering

Machine learning

SQL

Main component in Data science & Engineering

Compute – create the cluster

Workflows – create the jobs and schedule it

Workspace - code files present in workspace, we can create a notebook

Job cluster – sperate from other clusers, it is mainly for job schedule. automate the jobs,

We can work Python or Pyspark

Integrate the Blodge store with databricks

DBFS – uploaded

Blod\_Input = spark.read.option(“Header”, True).csv(“/mnt/demo/Sale\_data.csv”)

**Azure synapsis main components**

Develop 🡪 create a notebook

Integrator 🡪 create a pipeline

Monitor 🡪 monitor

Manage 🡪 create link service

Add the github using Repo

Use case

Use Databricks to analyze historical financial data, market trends, and customer behavior to assess and manage financial risks.

Workspace - create a notebook

To execute the jobs need to connect with the cluster

Magical command to add the comments

Connect the DAT folder

Databricks mapped to blob storage, and read the data from the stoge using pyspark,

ex : DF = spark. Read.option(“Header”, True).csv (“Demo/Mahes.CSV”)

hich service used in Azure for streaming data

ChatGPT

In Azure, the primary service used for streaming data is **Azure Stream Analytics**. This service enables real-time data processing and analytics on multiple streams of data from sources such as IoT devices, social media, applications, and more. Other related services include:

1. **Azure Event Hubs**: A big data streaming platform and event ingestion service.
2. **Azure IoT Hub**: Provides secure communication and data streaming for IoT devices.
3. **Azure Data Factory**: For orchestrating data movement and transformation, supporting streaming data integration.

These services can be used in combination to create robust streaming data solutions.

4o

where do you use the databricks ?

ChatGPT

Databricks is a cloud-based platform for big data analytics and machine learning. It is primarily used in a variety of data-related tasks and scenarios, particularly for organizations that deal with large datasets and want to leverage the power of Apache Spark for data processing. Here are some common use cases for Databricks:

1. **Data Transformation and ETL (Extract, Transform, Load):** Databricks is often used to preprocess and transform large volumes of raw data from various sources into a structured format that is suitable for analysis. This is a common task in data warehousing and data integration.
2. **Data Exploration and Analysis:** Data scientists and analysts use Databricks to explore and analyze data. The platform provides tools for running SQL queries, visualizing data, and performing statistical analysis. Apache Spark's distributed computing capabilities make it suitable for handling big data analysis.
3. **Machine Learning:** Databricks offers a collaborative environment for developing and deploying machine learning models. Data scientists can use it to build and train machine learning models at scale, taking advantage of distributed computing resources.
4. **Data Pipelines:** Organizations use Databricks to create data pipelines that automate the movement and transformation of data between different systems and storage solutions. This is essential for real-time analytics and reporting.
5. **Streaming Data Processing:** Databricks can process real-time streaming data using technologies like Apache Kafka and Apache Spark Streaming. This is useful for applications that require real-time insights from data streams.
6. **Predictive Analytics:** Databricks is used to build predictive analytics models that can forecast future trends and make data-driven decisions.
7. **Data Lakehouse:** Databricks Delta Lake, an extension of Apache Spark, is often used to create a data lakehouse architecture. It combines the best features of data lakes and data warehouses, allowing for both batch and real-time analytics on structured and semi-structured data.
8. **Collaboration:** Databricks provides collaborative features that enable data teams to work together on data projects. Multiple team members can collaborate on notebooks, share insights, and version control their work.
9. **Integration with Other Tools:** Databricks integrates with various data tools and platforms, including Apache Hadoop, AWS, Azure, and Google Cloud. This makes it easier for organizations to leverage their existing infrastructure and services.
10. **Security and Governance:** Databricks offers security and governance features to protect sensitive data and ensure compliance with data regulations. This is crucial for organizations that deal with sensitive or regulated data.

In summary, Databricks is used in a wide range of data-related scenarios, from data preprocessing and analysis to machine learning and real-time data processing. Its scalability, collaborative features, and integration capabilities make it a popular choice for organizations looking to derive insights from big data.



what is the difference between azure data factory and data bricks which one to use when ?

ChatGPT

Azure Data Factory and Databricks are both cloud-based data integration and processing services provided by Microsoft Azure, but they serve different purposes and are used in different scenarios. Here's a comparison of Azure Data Factory and Databricks and when to use each:

**Azure Data Factory:**

1. **Purpose:** Azure Data Factory is primarily an ETL (Extract, Transform, Load) and data integration service. It is designed for orchestrating and automating data workflows, moving data between different data stores, and transforming data at scale.
2. **Data Movement:** Data Factory is excellent for moving data from source systems to target systems, whether they are on-premises or in the cloud. It supports a wide range of connectors to various data sources and destinations.
3. **ETL and Data Transformation:** It provides a visual interface for building data pipelines that can transform data using data flows, mapping data transformations, and executing SQL scripts.
4. **Data Orchestration:** Azure Data Factory is great for scheduling and orchestrating data pipelines. You can set up triggers and dependencies to ensure that data workflows run at the right time and in the correct sequence.
5. **Data Integration:** Data Factory can integrate with various Azure services and third-party tools, making it suitable for integrating data from multiple sources into a data warehouse or data lake.

**Databricks:**

1. **Purpose:** Databricks is primarily a big data analytics and machine learning platform built on Apache Spark. It is designed for data exploration, advanced analytics, machine learning, and real-time data processing.
2. **Data Processing and Analytics:** Databricks excels at processing and analyzing large volumes of data using Apache Spark. It provides an interactive workspace for data scientists and analysts to run Spark jobs, write code in languages like Python and Scala, and build machine learning models.
3. **Real-time and Batch Processing:** Databricks can handle both real-time stream processing and batch processing of data. It's suitable for use cases that require near-real-time insights and processing of large datasets.
4. **Machine Learning:** Databricks has integrated machine learning libraries and capabilities, making it a powerful platform for developing and deploying machine learning models at scale.
5. **Data Exploration:** Data engineers and data scientists use Databricks for data exploration, visualization, and ad-hoc analysis. It provides collaborative notebooks for these tasks.

**When to Use Each:**

1. **Use Azure Data Factory when:**
   * You need to automate data movement and transformation workflows.
   * You want to orchestrate ETL processes involving various data sources and destinations.
   * Your primary goal is data integration and data pipeline orchestration.
2. **Use Databricks when:**
   * You need to perform advanced analytics, data exploration, and machine learning on large datasets.
   * You want to work with real-time data processing.
   * You require a collaborative environment for data scientists and analysts to build and deploy machine learning models.
   * Your primary goal is big data analytics and data science.

In many cases, these services can complement each other. For example, you might use Azure Data Factory to move and preprocess data before using Databricks for in-depth analysis and machine learning. The choice between Azure Data Factory and Databricks depends on your specific data integration and analytics needs.

Top of Form

Regenerate

hat are most commonly used services in azure ? which on to use when

ChatGPT

Microsoft Azure offers a wide range of cloud services and products, making it a versatile platform for various business needs. The choice of services depends on your specific use case and requirements. Here are some of the most commonly used services in Azure and when to use them:

1. **Azure Virtual Machines (VMs):**
   * Use when you need scalable, on-demand virtualized computing resources.
   * Ideal for running Windows or Linux-based applications and services.
2. **Azure App Service:**
   * Use for building and hosting web applications and APIs.
   * Supports various programming languages and frameworks.
3. **Azure SQL Database:**
   * Use when you need a fully managed relational database service.
   * Ideal for hosting SQL Server-based applications and data.
4. **Azure Blob Storage:**
   * Use for scalable object storage, including unstructured data, files, and backups.
   * Ideal for data storage and distribution.
5. **Azure Functions:**
   * Use for serverless computing to run code in response to events.
   * Ideal for microservices and event-driven architectures.
6. **Azure Kubernetes Service (AKS):**
   * Use for deploying, managing, and scaling containerized applications using Kubernetes.
   * Ideal for container orchestration and microservices.
7. **Azure Cosmos DB:**
   * Use when you need a globally distributed, NoSQL database with low-latency access.
   * Ideal for modern, highly available applications.
8. **Azure Active Directory (Azure AD):**
   * Use for identity and access management, including single sign-on (SSO) and multi-factor authentication (MFA).
   * Ideal for securing applications and resources.
9. **Azure Key Vault:**
   * Use for secure management of keys, secrets, and certificates.
   * Ideal for protecting sensitive information and managing encryption keys.
10. **Azure DevOps Services:**
    * Use for planning, developing, testing, and deploying applications.
    * Ideal for DevOps and continuous integration/continuous deployment (CI/CD) pipelines.
11. **Azure Logic Apps:**
    * Use for creating workflows and integrating services and systems.
    * Ideal for automating business processes.
12. **Azure Data Factory:**
    * Use for building, scheduling, and orchestrating data pipelines.
    * Ideal for data integration and ETL (Extract, Transform, Load) processes.
13. **Azure Databricks:**
    * Use for big data analytics, machine learning, and data exploration.
    * Ideal for data scientists and analysts working with large datasets.
14. **Azure Functions:**
    * Use for serverless computing to run code in response to events.
    * Ideal for microservices and event-driven architectures.
15. **Azure Synapse Analytics (formerly SQL Data Warehouse):**
    * Use for enterprise data warehousing and analytics.
    * Ideal for processing and analyzing large volumes of data.
16. **Azure IoT Hub:**
    * Use for connecting, monitoring, and managing IoT devices.
    * Ideal for building IoT solutions.
17. **Azure Cognitive Services:**
    * Use for adding AI and machine learning capabilities to applications.
    * Ideal for image recognition, natural language processing, and more.

When choosing Azure services, consider factors like scalability, cost, security, and integration with your existing systems and tools. Azure offers a wide array of services to cater to various business needs, so it's essential to carefully evaluate your requirements and select the services that best fit your use case. Additionally, Azure's hybrid capabilities allow you to seamlessly integrate on-premises and cloud solutions to create hybrid cloud architectures when necessary.

Top of Form

Regenerate

what are most commonly used services in AWS ? which on to use when

ChatGPT

Amazon Web Services (AWS) provides a vast array of cloud services to cater to various business needs. The choice of AWS services depends on your specific use case and requirements. Here are some of the most commonly used services in AWS and when to use them:

1. **Amazon EC2 (Elastic Compute Cloud):**
   * Use when you need scalable, on-demand virtualized computing resources.
   * Ideal for running virtual machines (VMs) to host applications and services.
2. **Amazon S3 (Simple Storage Service):**
   * Use for scalable object storage, including data, files, backups, and multimedia content.
   * Ideal for data storage and distribution.
3. **Amazon RDS (Relational Database Service):**
   * Use when you need a fully managed relational database service.
   * Ideal for hosting databases like MySQL, PostgreSQL, SQL Server, and Oracle.
4. **Amazon Lambda:**
   * Use for serverless computing to run code in response to events.
   * Ideal for microservices, event-driven architectures, and automation.
5. **Amazon ECS (Elastic Container Service):**
   * Use for deploying and managing containers using Docker.
   * Ideal for container orchestration and microservices.
6. **Amazon DynamoDB:**
   * Use when you need a NoSQL database that provides fast and scalable access to data.
   * Ideal for modern, high-traffic applications.
7. **Amazon SNS (Simple Notification Service):**
   * Use for sending notifications and messages to distributed systems.
   * Ideal for event-driven communication in applications.
8. **Amazon IAM (Identity and Access Management):**
   * Use for managing user and service access to AWS resources.
   * Ideal for securing AWS services and resources.
9. **Amazon VPC (Virtual Private Cloud):**
   * Use to create isolated network environments in the AWS cloud.
   * Ideal for network security and segmentation.
10. **Amazon Route 53:**
    * Use for domain name system (DNS) and domain registration services.
    * Ideal for routing internet traffic to AWS resources.
11. **Amazon SQS (Simple Queue Service):**
    * Use for building scalable and decoupled distributed applications.
    * Ideal for managing message queues.
12. **Amazon Kinesis:**
    * Use for real-time streaming and analytics of data.
    * Ideal for applications that require real-time processing of data streams.
13. **Amazon RDS (Relational Database Service):**
    * Use when you need a fully managed relational database service.
    * Ideal for hosting databases like MySQL, PostgreSQL, SQL Server, and Oracle.
14. **Amazon EKS (Elastic Kubernetes Service):**
    * Use for deploying, managing, and scaling containerized applications using Kubernetes.
    * Ideal for container orchestration and microservices.
15. **Amazon Redshift:**
    * Use for data warehousing and analytics with high-performance querying.
    * Ideal for processing and analyzing large volumes of data.
16. **Amazon Elasticache:**
    * Use for in-memory data caching to speed up data retrieval.
    * Ideal for improving the performance of web applications.
17. **AWS Glue:**
    * Use for ETL (Extract, Transform, Load) and data integration.
    * Ideal for building data pipelines and data transformation.
18. **Amazon Polly:**
    * Use for text-to-speech (TTS) conversion in applications.
    * Ideal for adding voice capabilities to applications.
19. **Amazon Comprehend:**
    * Use for natural language processing (NLP) and sentiment analysis.
    * Ideal for analyzing text data.
20. **Amazon SageMaker:**
    * Use for building, training, and deploying machine learning models.
    * Ideal for data scientists and developers working on ML projects.

When selecting AWS services, consider factors like scalability, cost, security, and integration with your existing systems and tools. AWS provides a wide range of services to address various business needs, so it's important to evaluate your requirements and choose the services that best suit your use case. Additionally, AWS offers hybrid cloud solutions for integrating on-premises and cloud resources to create hybrid architectures when needed.

Top of Form

**Azure Service Bus: equallent AWS SQS**

Azure Stream Analytics is the equivalent service to Amazon Kinesis in Microsoft Azure.

Azure Synapse Analytics is the equivalent service to Amazon Redshift in Microsoft Azure.

Azure App Service and AWS Elastic Beanstalk:

Azure DNS is the equivalent service to Amazon Route 53 in Microsoft Azure.

where do you use the azure databricks in banking industory

ChatGPT

Azure Databricks can be a valuable tool for the banking industry, providing a platform for data analytics, machine learning, and data engineering. Here are several use cases and scenarios where Azure Databricks can be applied in the banking sector:

1. **Risk Assessment and Management:**
   * Use Databricks to analyze historical financial data, market trends, and customer behavior to assess and manage financial risks. Machine learning models can help predict credit risk, fraud detection, and market volatility.
2. **Customer Analytics:**
   * Utilize Databricks to analyze customer data to gain insights into behavior and preferences. Personalized banking recommendations and targeted marketing campaigns can be developed based on these insights.
3. **Fraud Detection:**
   * Databricks can be used to build real-time fraud detection systems by analyzing transaction data and detecting unusual patterns or anomalies that may indicate fraudulent activity.
4. **Anti-Money Laundering (AML):**
   * Apply Databricks for AML compliance by analyzing large volumes of transaction data and identifying suspicious activities that may be indicative of money laundering.
5. **Regulatory Compliance:**
   * Databricks can help streamline compliance efforts by automating data aggregation and reporting tasks required by regulatory bodies. This ensures that banks are compliant with regulations such as Basel III, GDPR, or Dodd-Frank.
6. **Customer Churn Prediction:**
   * Use machine learning models in Databricks to predict customer churn. By identifying customers at risk of leaving, banks can take proactive measures to retain them.
7. **Credit Scoring:**
   * Employ Databricks to develop credit scoring models that assess the creditworthiness of applicants based on various factors. This aids in the loan approval process.
8. **Operational Efficiency:**
   * Optimize banking operations by using Databricks for data engineering tasks. Data can be cleaned, transformed, and loaded into data warehouses efficiently, making it accessible for analysis.
9. **Real-time Analytics:**
   * Databricks supports real-time stream processing. Banks can use it to monitor transactions and events in real time, enabling timely responses to issues or opportunities.
10. **Market Data Analysis:**
    * Analyze market data feeds and news in real time to inform trading decisions and investment strategies.
11. **Wealth Management:**
    * For wealth management services, use Databricks to provide personalized investment recommendations based on customer financial goals and risk tolerance.
12. **Cybersecurity:**
    * Enhance cybersecurity by using Databricks to analyze network and security data for threats and vulnerabilities.
13. **Asset and Portfolio Management:**
    * Analyze asset and portfolio performance to make informed investment decisions, optimize portfolios, and maximize returns.
14. **Cost Optimization:**
    * Databricks can help banks optimize costs by analyzing operational data and identifying areas for cost reduction and process improvement.

In summary, Azure Databricks offers a versatile platform for data analytics and machine learning, making it a valuable asset for banks seeking to enhance their operations, improve customer experiences, and meet regulatory requirements. Its ability to process large volumes of data, support real-time analytics, and facilitate machine learning makes it a powerful tool for addressing various challenges and opportunities in the banking industry.

Azure azure databricks interview question and answers for 15 years experience

ChatGPT

Here are some interview questions and answers for a candidate with 15 years of experience in Azure Databricks:

1. **What is Azure Databricks, and how does it differ from Apache Spark?**

*Answer:* Azure Databricks is a cloud-based big data analytics platform that integrates with Apache Spark. It provides a collaborative environment for data scientists, data engineers, and analysts to work together on big data and machine learning projects. The main difference from Apache Spark is the managed and fully integrated nature of Azure Databricks within the Azure ecosystem.

1. **Explain the architecture of Azure Databricks.**

*Answer:* Azure Databricks has a distributed architecture. It consists of a control plane that manages clusters and notebooks, and a data plane that stores data and runs computations. The control plane includes the Azure Databricks Workspace for collaboration and job scheduling. The data plane includes data storage, often using Azure Data Lake Storage or Azure Blob Storage, and clusters for distributed data processing.

1. **What are some common use cases for Azure Databricks?**

*Answer:* Azure Databricks is used for a variety of data processing and analysis tasks, including real-time analytics, ETL (Extract, Transform, Load) processes, machine learning, data exploration, and data warehousing. It's suitable for industries like finance, healthcare, retail, and more.

1. **How can you optimize the performance of Azure Databricks clusters?**

*Answer:* Performance optimization in Azure Databricks can be achieved by tuning cluster configuration, leveraging autoscaling, caching intermediate data, optimizing your Spark jobs, and using efficient data storage formats like Parquet. Additionally, you can monitor and analyze cluster performance using built-in tools.

1. **What is Delta Lake, and how does it enhance data reliability in Azure Databricks?**

*Answer:* Delta Lake is a storage layer that brings ACID (Atomicity, Consistency, Isolation, Durability) transactions to data lakes. It enhances data reliability in Azure Databricks by allowing you to perform operations like insert, update, and delete on data stored in data lakes, ensuring data consistency and reliability.

1. **Explain the different security features available in Azure Databricks.**

*Answer:* Azure Databricks offers various security features such as Azure Active Directory integration, role-based access control (RBAC), workspace and cluster access controls, encryption at rest and in transit, audit logging, and network security. These features help secure data and access within the platform.

1. **What is the difference between Databricks Runtime and Databricks Runtime ML in Azure Databricks?**

*Answer:* Databricks Runtime is the core runtime environment for processing data and running Spark workloads. Databricks Runtime ML is an extension that includes additional libraries and tools for machine learning. It comes with pre-installed ML frameworks like TensorFlow and scikit-learn.

1. **How can you monitor and troubleshoot performance issues in Azure Databricks?**

*Answer:* Azure Databricks provides built-in monitoring and logging capabilities. You can use tools like Azure Monitor, Azure Log Analytics, and Databricks Cluster Logs to monitor and troubleshoot performance issues. You can also leverage third-party monitoring and alerting solutions.

1. **What is the difference between a job and a notebook in Azure Databricks?**

*Answer:* A notebook in Azure Databricks is an interactive environment for developing and running code, whereas a job is a way to schedule the execution of a notebook or a JAR file at a specified time or on a trigger. Jobs are used for automation and productionizing code developed in notebooks.

1. **Explain the process of migrating on-premises Spark workloads to Azure Databricks.**

*Answer:* Migrating on-premises Spark workloads to Azure Databricks involves tasks like data migration, rewriting code to work with Azure Databricks, configuring clusters, and testing. Azure offers tools like Azure Data Factory and Azure Database Migration Service to assist in this process.

Remember that with 15 years of experience, interviewers may also ask about your experience with specific projects, challenges you've faced, and how you've contributed to the success of previous teams and organizations. Be prepared to provide examples of your real-world experiences with Azure Databricks to demonstrate your expertise.

Top of Form

Regenerate