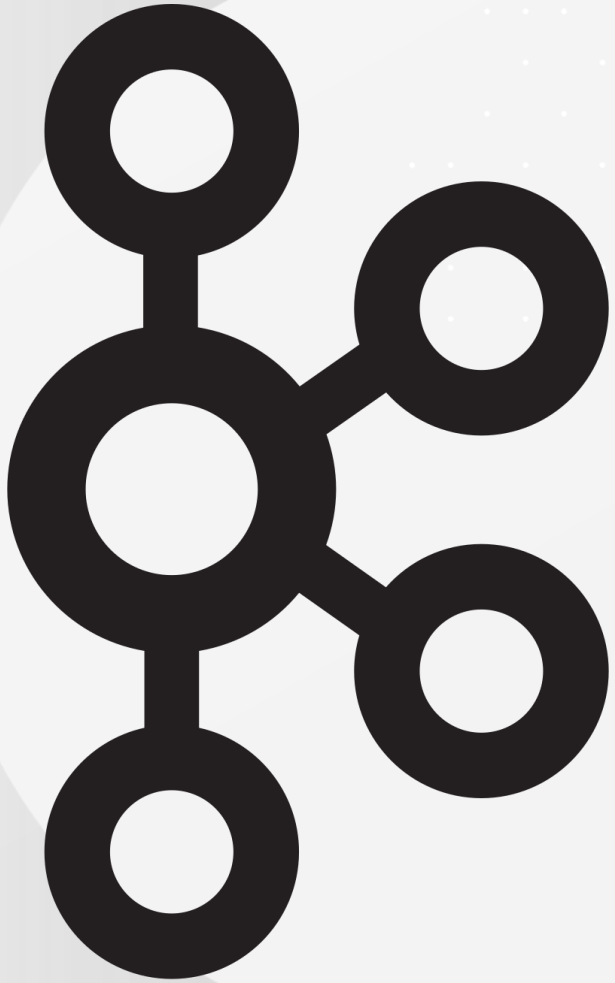
BPP

Advanced Kafka scenarios

Welcome to today’s webinar.

According to Confluent's surveys, **81% of companies use Kafka for data pipelines**, and **66% use it for stream processing.**



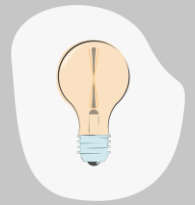
Real-world case study

Audi’s use of Kafka for predictive maintenance

- Audi uses Apache Kafka for predictive maintenance.
- Monitors vehicle health and predicts failures in real-time.

Key benefits:

- Real-Time Data Processing
- Scalability
- Reliability

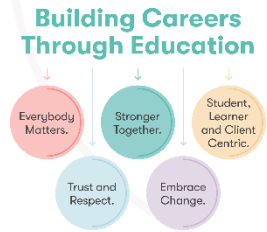


Impact on business:

- Reduces maintenance costs.
- Improves vehicle reliability.
- Enhances customer satisfaction by preventing breakdowns.

The Audi R8

Knowledge check discussion



Imagine you are a data engineer at an automotive company that uses Kafka to monitor vehicle health data in real-time. Recently, you noticed that some sensor data is delayed, leading to inaccurate predictive maintenance alerts.

What steps would you take to identify and resolve the issue of delayed sensor data in your Kafka system?

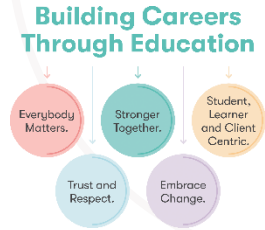
Consider the following:

- Identifying symptoms
- Analysing potential causes
- Implementing solutions
- Preventative measures

Submit your responses to the chat or turn on your microphone!

BPP

Session aim and objectives



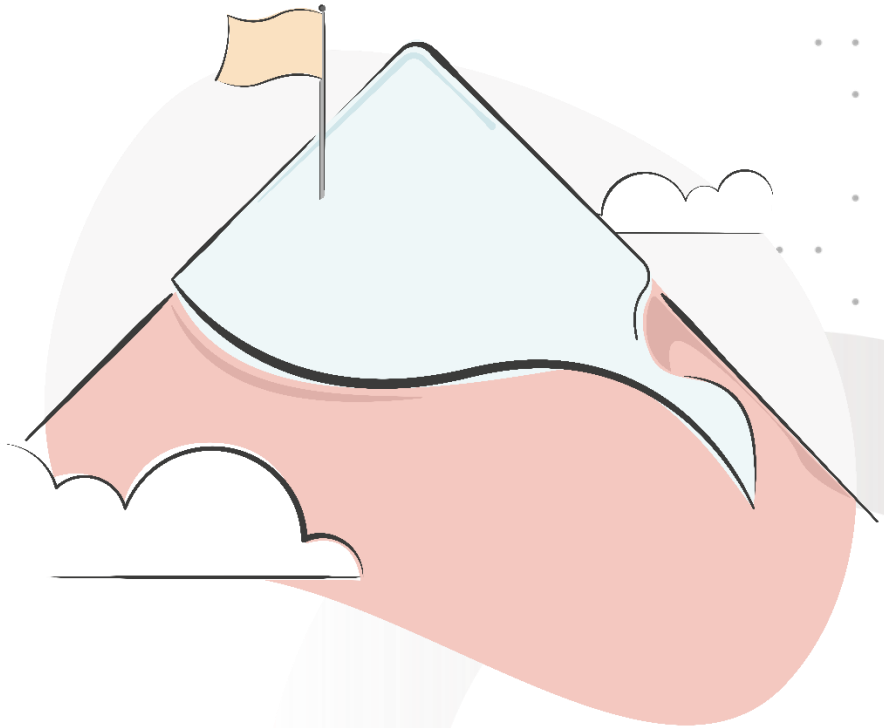
By the end of this session, you should be able to:

1. Compare Azure Event Hubs and Event Grid with Kafka, including motivations and contrasts with AWS and GCP services.

2. Implement SSL, SASL, and ACL for secure Kafka communication.

3. Develop troubleshooting skills and conduct root-cause analysis for Kafka-related issues.

4. Apply structured root-cause analysis techniques in practical scenarios.



BPP

Cloud-Based Event Streaming Services

Event streaming in big data

Big Data Landscape:

- Real-time data processing is essential.
- Event streaming platforms are crucial.

Apache Kafka:

- Leader in high-throughput streaming.
- Widely adopted open-source solution.

Cloud Providers:

- Azure, AWS, GCP offer managed services.
- Benefits: managed infrastructure, seamless integration, scalability.

Azure, AWS, and Google Cloud
Image source: LinkedIn

BPP

Azure Event Hubs and Event Grid

Big Data and Real-Time Processing

Comparison	Azure Event Hubs	Azure Event Grid
Definition	Managed, real-time data ingestion service.	Managed event routing service.
Use Cases	Telemetry ingestion, log collection, real-time analytics, big data pipelines	Serverless applications, workflow automation, event integration.
Key Features	• Supports AMQP 1.0 and Kafka protocols. • Partitioning and offset management. • Capture feature for storing data in Azure Blob Storage or Data Lake.	• Low latency event delivery. • Advanced filtering and routing. • Supports custom events and integrates with Azure services.

Event Grid vs Event Hub logos
Image source: Azure Lessons

Comparing Azure Event Hubs with Kafka

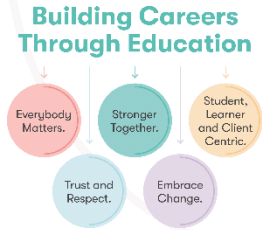
Azure Event Hubs vs Kafka

Feature	Kafka	Azure Event Hubs
Management	Self-managed clusters, infrastructure provisioning, and maintenance	Fully managed service, abstracts cluster management complexities
Protocol support	Uses its own protocol; extensive client libraries	Kafka-compatible endpoint; no client code changes needed
Ecosystem and integration	Rich ecosystem with many connectors and community support	Deep integration with Azure services; may have limitations outside Azure
Scalability	Highly scalable with fine-grained perfor mance control	Scales automatically with less control over infrastructure

VS

BPP

AWS and GCP comparisons



Services and feature comparisons…

Service	Definition	Features
AWS Kinesis	Scalable and durable real-time data streaming service	Real-time data ingestion; custom development needed for complex processing
AWS SNS	Pub-sub messaging service	Simplifies message distribution to multiple subscribers
AWS SQS	Message queuing service	Decouples components, ensuring reliable message delivery
GCP Pub/Sub	Global, scalable messaging service	At-least-once delivery, message ordering, integration with GCP services

VS

Comparing AWS and GCP services with Apache Kafka

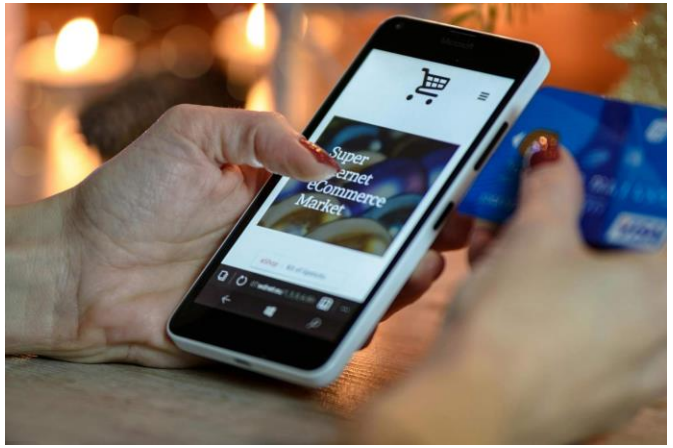
BPP

Event streaming and Kafka

Industry use cases

Financial Services:

- **Use Case:** Real-time fraud detection. •
- **Details:** Process transaction data in real-time to identify and prevent fraud.•



E-commerce:

Use Case: Personalised recommendations.
Details: Analyse user behavior in real-time for personalised product recommendations.

IoT and Manufacturing:

- **Use Case:** Predictive maintenance.
- **Details:** Monitor IoT sensor data to predict failures and schedule maintenance proactively.

How could implementing event streaming solutions like Kafka address challenges in your organisation?

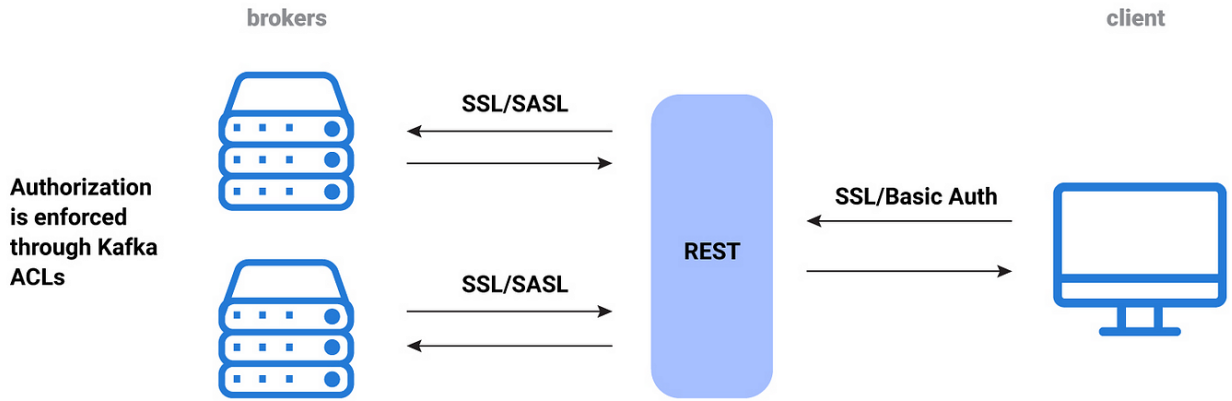
Overview of SSL for Kafka

Encryption and integrity for data in transit

- **Definition:** SSL (Secure Sockets Layer) and its successor TLS (Transport Layer Security) provide encryption and integrity for data in transit.

Importance of SSL:

- **Data Protection:** Encrypts data, preventing eavesdropping and attacks.
- **Compliance:** Meets GDPR, HIPAA requirements for data protection.
- **Internal Security:** Prevents insider threats by encrypting internal communications.



Kafka with TLS/SSL
Image source: Medium

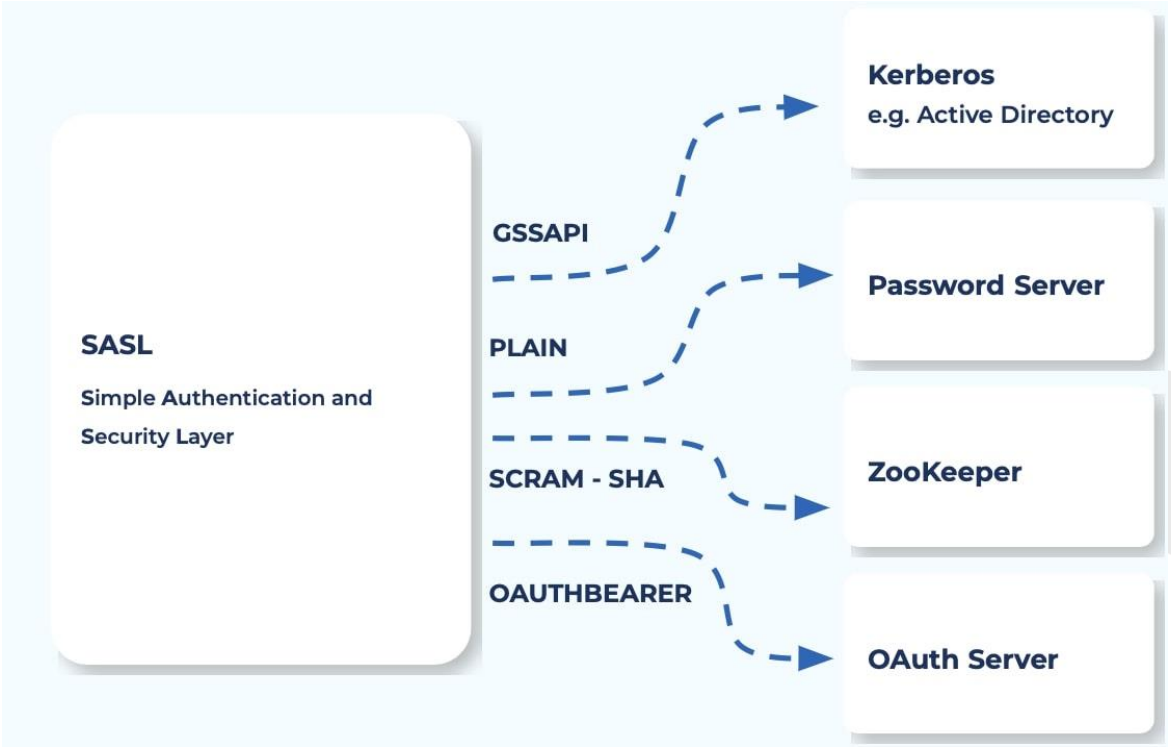
Overview of SASL mechanisms

Ensuring Kafka communication

- **Definition:** SASL (Simple Authentication and Security Layer) provides a framework for authentication and data security in Kafka.

Supported Mechanisms:

- **GSSAPI (Kerberos):** Strong authentication using Kerberos.
- **PLAIN:** Simple username and password authentication.
- **SCRAM (Salted Challenge Response Authentication Mechanism):** Enhanced security over PLAIN.
- **OAuth:** Token-based authentication.



Kafka security protocols
Image source: Confluent Developer

BPP

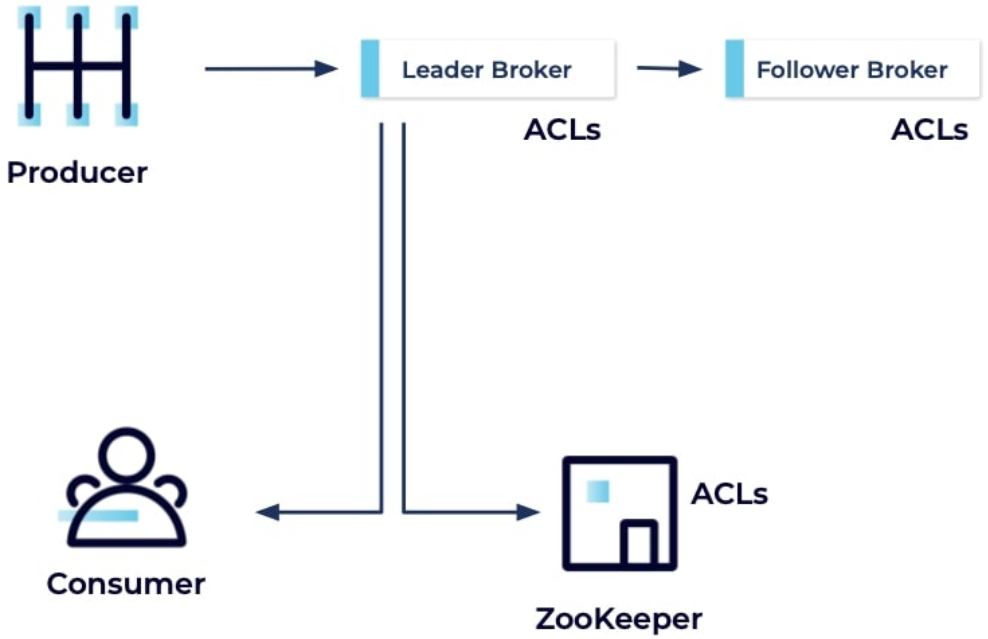
Overview of ACL mechanisms

Controlling Kafka authentication

- **Definition:** ACLs (Access Control Lists) authorise actions performed by authenticated users, controlling who can read, write, or administer Kafka resources.

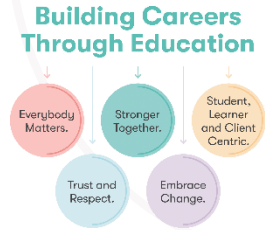
Purpose:

- **Authorisation:** Define permissions for users and roles.
- **Control:** Manage who can read, write, or administer Kafka topics and resources.



Kafka Authorisation and Access Control Lists (ACL)
Image source: Confluent Developer

Troubleshooting Kafka issues



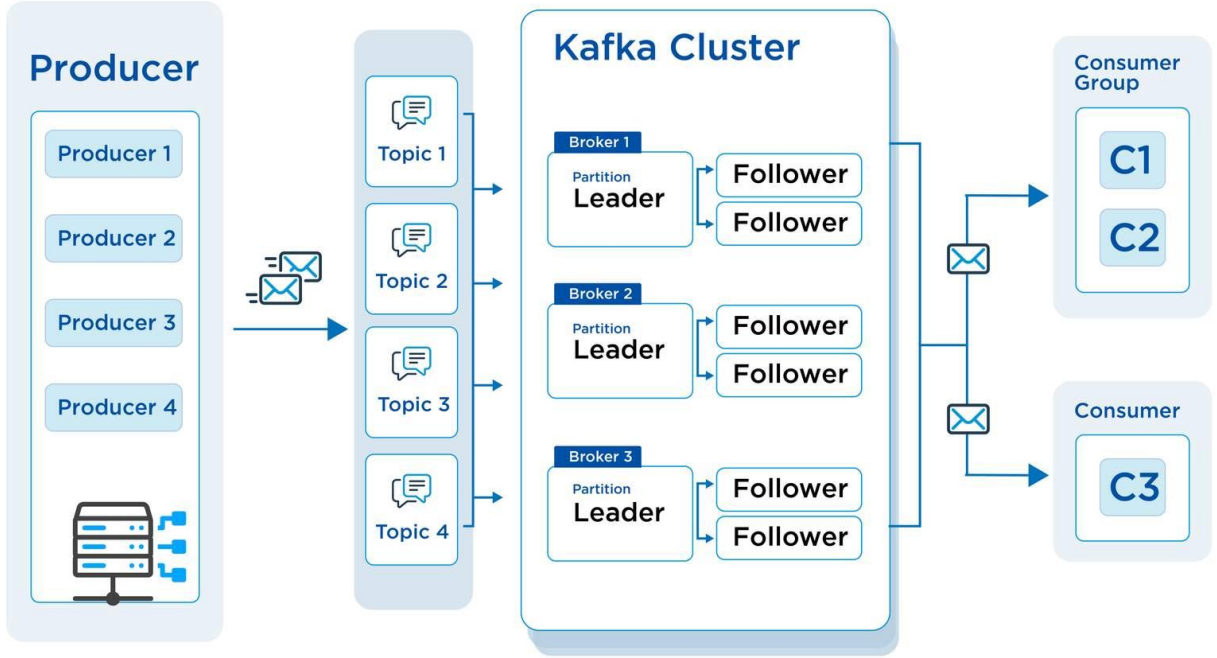
Common issues, symptoms, and potential causes

Kafka Cluster:

- Multiple brokers manage storage and retrieval
- Zookeeper ensures consistency and reliability
- Scales horizontally by adding brokers

Data Consistency:

- Replication: multiple replicas across brokers
- Leader and follower brokers for each partition
- Strong consistency: all replicas acknowledge messages
- Ensures availability and consistency even if a broker fails



An illustration of the Kafka ecosystem
Image source: LinkedIn

BPP

Troubleshooting Kafka issues

Common issues, symptoms, and potential causes

Issue	Symptoms	Causes
Consumer lag	Delayed message processing, increasing offset lag	Insufficient consumer capacity, slow processing logic, network bottlenecks
Broker failure	Unresponsive brokers, errors in producer or consumer clients	Hardware failures, JVM crashes, disk full errors
Under-replicated partitions	Partitions with less than configured replication factor, alerts in monitoring systems	Broker outages, network issues preventing replication
Message loss	Missing messages, discrepancies in da ta processing	Incorrect producer configurations, early data deletion, consumer offset mismanagement
High latency	Increased time for producers to send m essages, delays in consumers receivin g messages	Network congestion, overloaded brokers, slow disk I/O

Structured root cause analysis techniques

5 whys, Fishbone diagram, and fault tree analysis

The 5 Whys Technique:

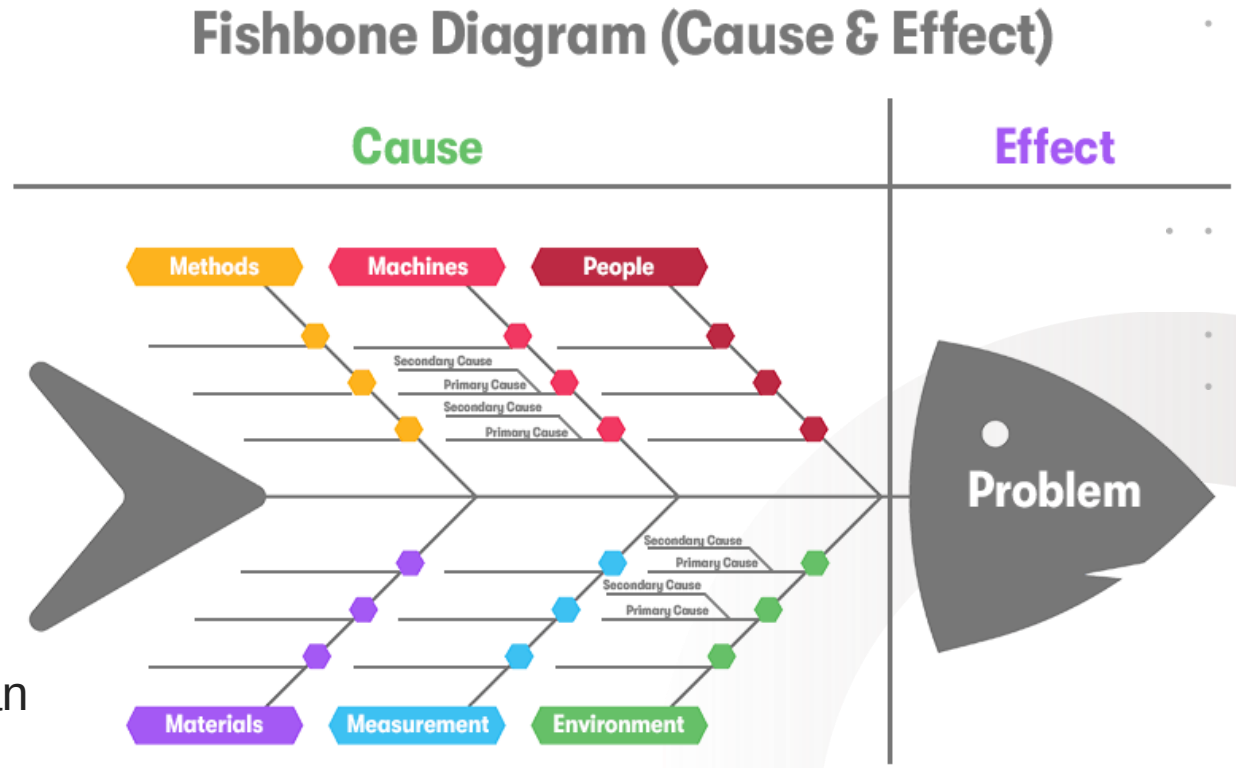
- **Method:** Ask "Why?" iteratively to uncover the root cause.

Fishbone Diagram (Ishikawa Diagram):

- **Method:** Visual tool categorising causes under headings like People, Processes, Technology, Environment.

Fault Tree Analysis:

- **Method:** Top-down approach using Boolean logic to model failure pathways.



A fishbone diagram
Image source: PurpleGriffon

Time for the practical lab!

Your tutor will provide guidance as required…

Practical lab activities detailed in this document: Lab activities

Key Learning Summary

Here is a summary of the key learning points for this topic:

- Azure Event Hubs is a fully managed, real-time data ingestion service.
- Azure Event Grid is a fully managed event routing service with low latency.
- SSL/TLS encrypts data in transit for confidentiality and integrity.
- SASL provides authentication using mechanisms like Kerberos and SCRAM.
- ACLs control access to Kafka resources by defining user permissions.
- Common Kafka issues include consumer lag, broker failures, and high latency.
- The 5 Whys and Fishbone Diagram help identify root causes in troubleshooting.

BPP

Session aim and objectives

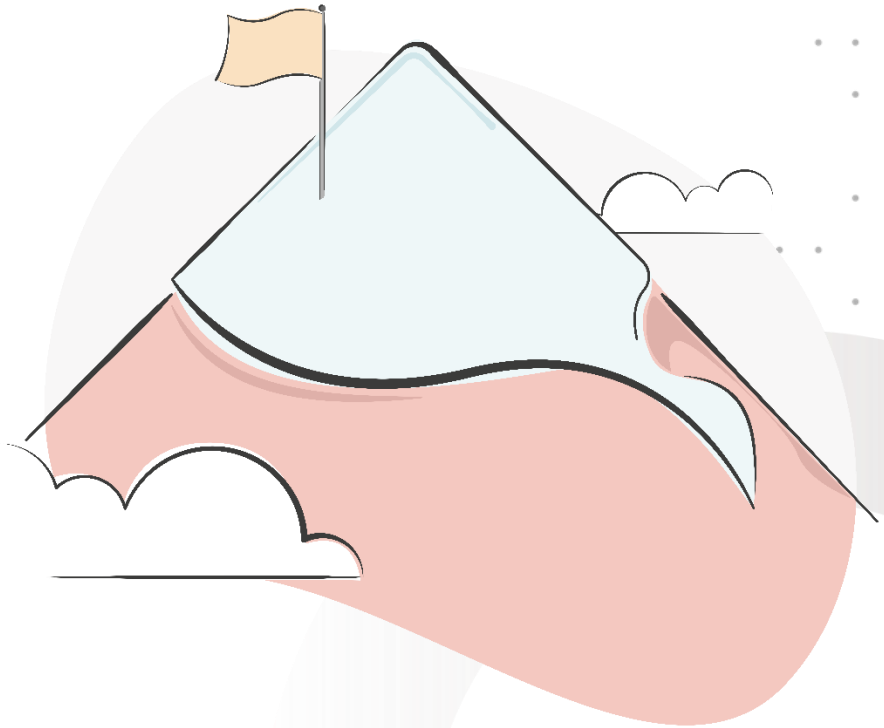
You should now be able to:

1. Compare Azure Event Hubs and Event Grid with Kafka, including motivations and contrasts with AWS and GCP services.

2. Implement SSL, SASL, and ACL for secure Kafka communication.

3. Develop troubleshooting skills and conduct root-cause analysis for Kafka-related issues.

4. Apply structured root-cause analysis techniques in practical scenarios.



BPP

Thank you

Do you have any questions, comments, or feedback?