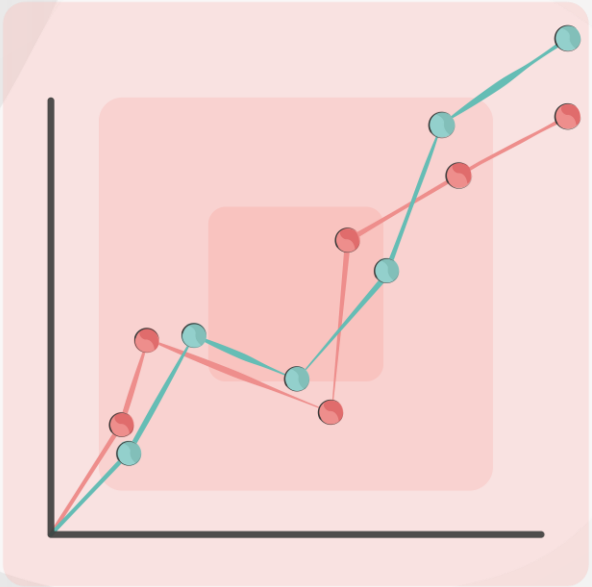
BPP

Monitoring an ingestion service and anomaly detection techniques

Welcome to today's webinar.



Case study application

Building Careers Through Education

Everybody Matters.

Stronger Together.

Student, Learner and Client Centric.

Trust and Respect.

Embrace Change.

Real-time anomaly detection in network traffic

Consider the follow use case…

- **The challenge:** High transaction volumes in financial institutions

- **Need:** To prevent system issues and detect fraud

- **Implementation**: Isolation Forest, time series models, Striim integration

The outcome:

- Reduced impact of issues, improved efficiency, enhanced security

Image source, Freepik.com, [Link]()

Knowledge check poll

Building Careers Through Education

Everybody Matters.

Stronger Together.

Student, Learner and Client Centric.

Trust and Respect.

Embrace Change.

In the case study of a financial institution using real-time anomaly detection.

Which of the following techniques was NOT mentioned as part of their implementation?

A. Isolation Forest

B. Time Series Models

C. Encryption

D. Striim Integration

Correct answer: C - While encryption is important for data security, it was not mentioned as part of the anomaly detection implementation in the case study.

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

Session aim and objectives

Building Careers Through Education

Everybody Matters.

Stronger Together.

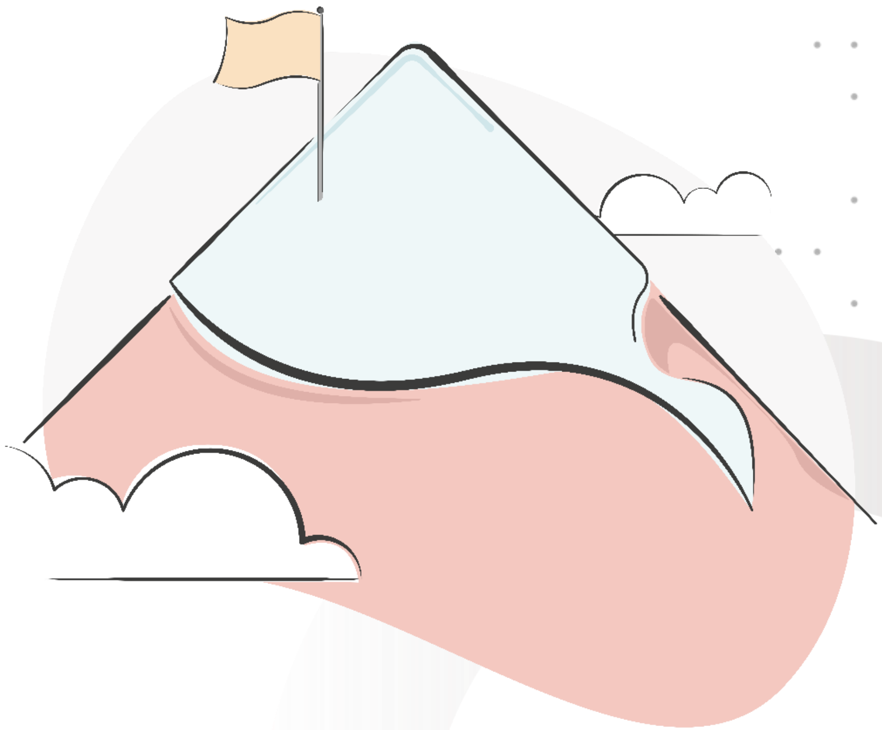
Student, Learner and Client Centric.

Trust and Respect.

Embrace Change.

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

- Automate monitoring processes for data ingestion services using industry-standard tools.
- Implement forecasting and anomaly detection techniques, including ARIMA, SARIMAX, and other methods.
- Integrate monitoring with incident management systems to enhance operational responsiveness.
- Address real-world use cases and typical ingestion issues in Kafka and cloud environments.



Monitoring in data engineering

Why is it important?

Monitoring data ingestion pipelines is essential to:

- Ensure Data Integrity
- Maintain Performance
- Enhance Reliability
- Support Compliance

Image source, Freepik.com, Link

Key concepts in monitoring:

- Metrics, logs, alerts, and dashboard

What do you know about these concepts?

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

BPP

Industry-standard monitoring tools

Prometheus and Grafan

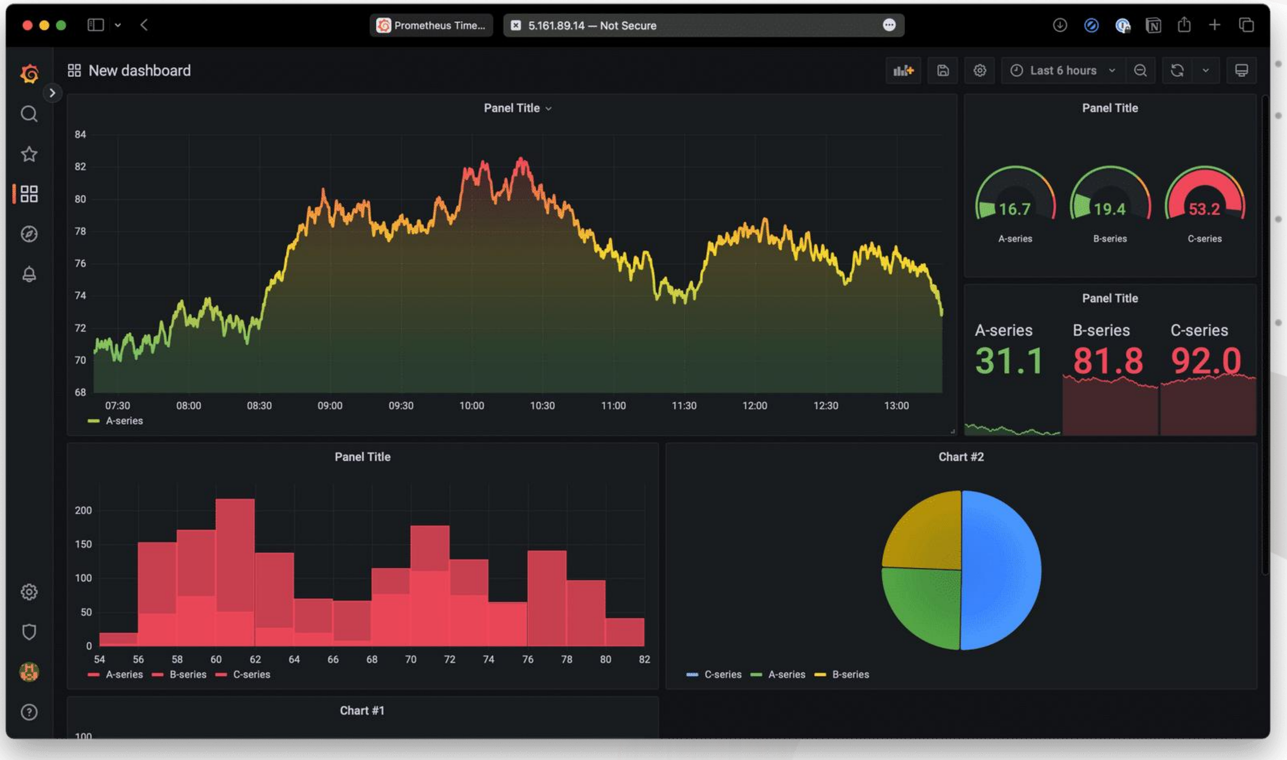
These tools are essential for monitoring and visualising data pipelines:

Prometheus:

- Open-source monitoring toolkit
- Collects metrics at intervals
- Time series data model
- PromQL for data analysis
- Alertmanager for alerts

Grafana:

- Open-source visualisation platform
- Supports multiple data sources
- Customisable dashboards
- Integrated alerting



A Prometheus dashboard

Building Careers Through Education

Everybody Matters.

Stronger Together.

Student, Learner and Client Centric.

Trust and Respect.

Embrace Change.

BPP

Monitoring Kafka with Prometheus and Grafana

How is this useful?

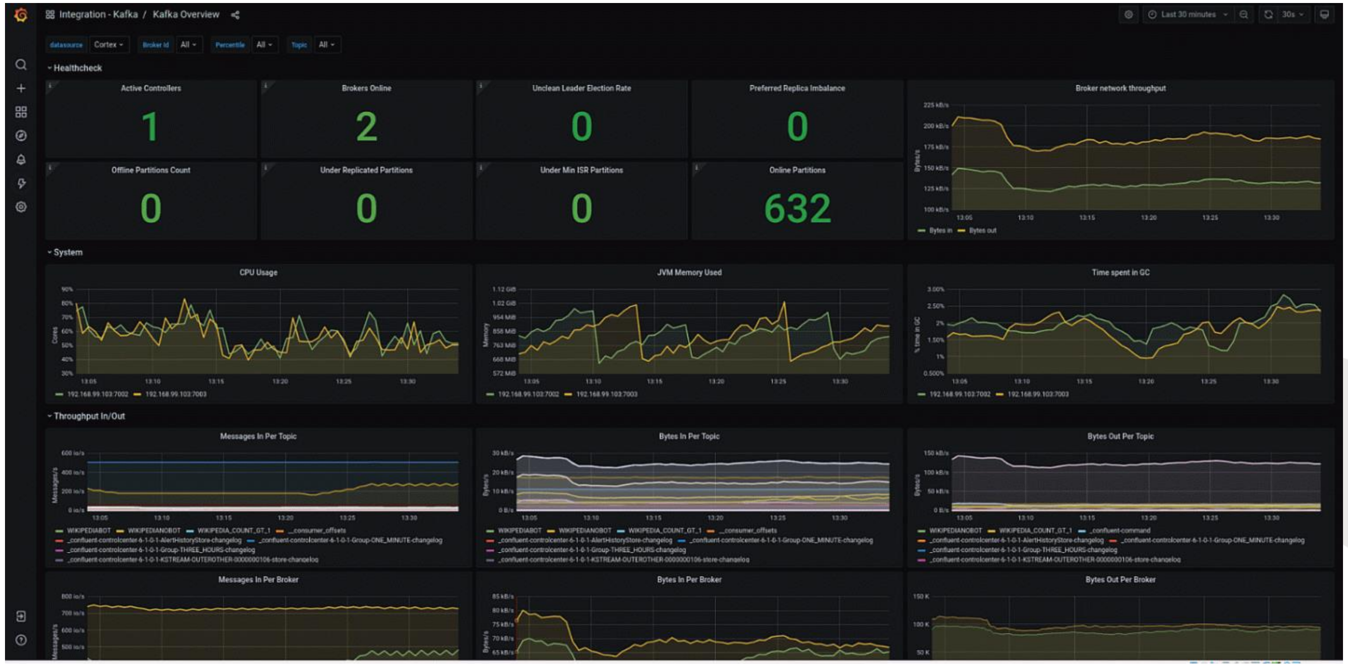
Ensures smooth operation of Kafka clusters and provides valuable insights into Kafka metrics.

Kafka Exporter:

- Purpose: Collects Kafka metrics and exposes them to Prometheus
- Installation: Download binary or build from source
- Configuration: Connect to Kafka cluster with appropriate flags

Key Metrics to Monitor:

- Broker Metrics
- Producer Metrics
- Consumer Metrics



An example of Apache Kafka ecosystem monitoring

Automating monitoring processes

Alerting with Prometheus Alertmanager

Beyond metrics collection and visualisation, it automates responses to potential issues.

Alertmanager Setup:

- Define alerting rules
- Configure notification channels

Example Alerting Rule:

- HighConsumerLag alert for Kafka
- Expression: kafka_consumer_lag > 10000
- Duration: 5 minutes
- Severity: Critical
- Annotations: Summary and description

```yaml
groups:
  - name: kafka_alerts
    rules:
      - alert: HighConsumerLag
        expr: kafka_consumer_lag > 10000
        for: 5m
        labels:
          severity: critical
        annotations:
          summary: "High Consumer Lag Detected"
          description: "Consumer lag is {{
$value }} for group {{ $labels.group }}"
```

An alerting rules example

Case study scenario

Monitoring a healthcare data pipeline

How would you approach this task…?

Implementation steps:

1. Deploy Prometheus and Grafana
2. Set Up Kafka Exporter
3. Create Dashboards
4. Define Alerting Rules
5. Integrate with Incident Management

Benefits:

- Improved patient care, enhanced data integrity, and operational efficiency

Image source: Actian.co, link

Understanding the need for forecasting

A critical component in managing data ingestion services

Importance of Forecasting:

- Predict future system behavior
- Plan capacity
- Prevent overloads
- Optimise costs

Application:

Forecasting can help retail chains to plan inventory levels, prevent stockouts and overstock situations, and optimise staffing.

The retailer John Lewis is the type of retail chain that might need to forecast demand

Understanding Time Series Forecasting

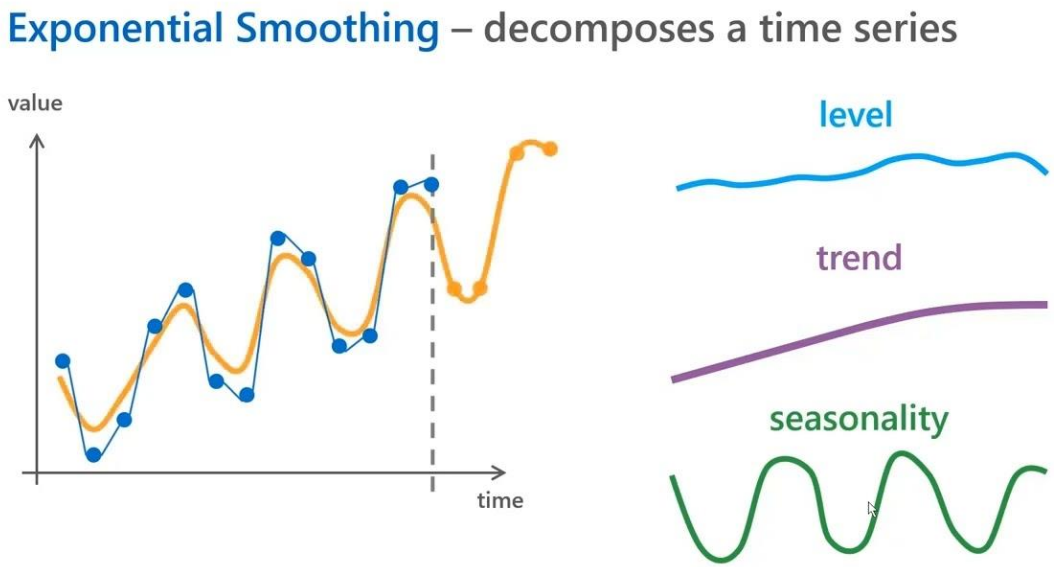
Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Analysing historical data to predict future values

Common Methods:

- **ARIMA:** AutoRegressive Integrated Moving Average
- **SARIMAX:** Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors
- Prophet
- Machine Learning Models (e.g., LSTM networks)



Time Series Forecasting image source, YouTube: Link

- **Level**: The baseline value of the series
- **Trend**: The direction and rate of change over time
- **Seasonality**: Regular, repeating patterns or cycles

Anomaly detection techniques

Statistical, machine learning and time series

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.

Statistical Methods:

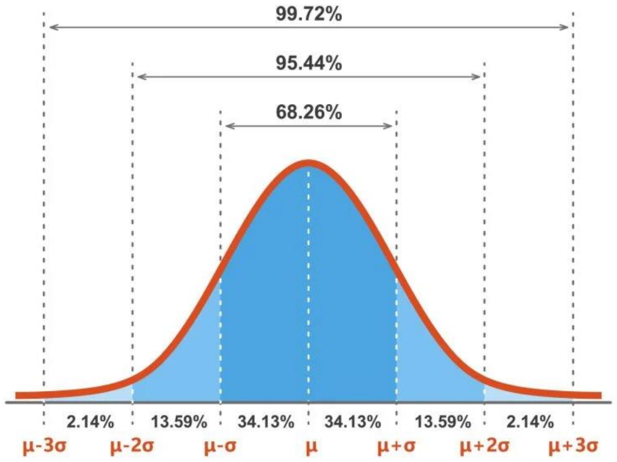
- Z-Score Analysis
- Seasonal Hybrid ESD

Machine Learning Models:

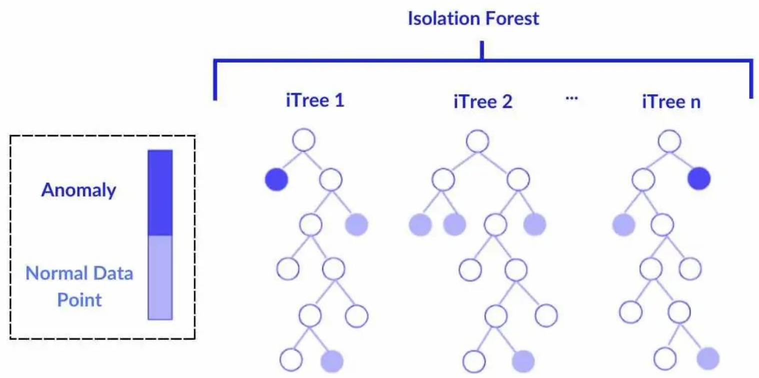
- Isolation Forest
- One-Class SVM

Time Series Models:

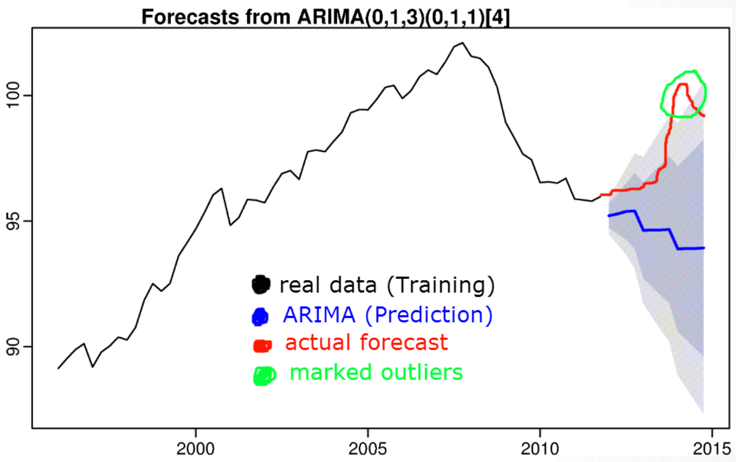
- ARIMA/SARIMAX Residual Analysis
- Prophet



An illustration of Z-score, link



An illustration of Isolation forest, link



An illustration ofARIMA, link

BPP

Time for the practical lab!

Your tutor will provide guidance as required…

Building Careers Through Education

Everybody Matters.

Stronger Together.

Student, Learner and Client Centric.

Trust and Respect.

Embrace Change.

Practical lab activities detailed in this document: Lab activities

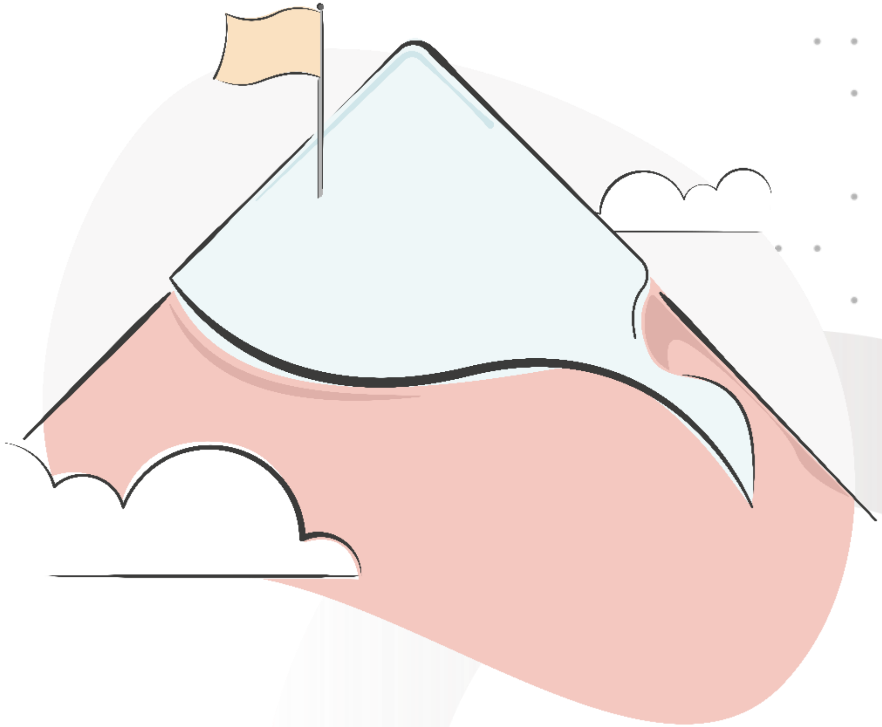
BPP

Session aim and objectives

Building Careers Through Education
Everybody Matters.
Stronger Together.
Student, Learner and Client Centric.
Trust and Respect.
Embrace Change.

You should now be able to:

- Automate monitoring processes for data ingestion services using industry-standard tools.
- Implement forecasting and anomaly detection techniques, including ARIMA, SARIMAX, and other methods.
- Integrate monitoring with incident management systems to enhance operational responsiveness.
- Address real-world use cases and typical ingestion issues in Kafka and cloud environments.



Key Learning Summary

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

- **Prometheus and Grafana** are essential tools for monitoring data ingestion pipelines, providing metrics collection, alerting, and visualisation to ensure system reliability and performance.

- **Forecasting techniques** like ARIMA and SARIMAX are crucial for predicting future data ingestion rates, allowing for effective capacity planning and preventing system overloads.

- **Anomaly detection methods**, such as Isolation Forest and Seasonal Hybrid ESD, are vital for identifying unusual patterns in data streams

- **Integrating Prometheus Alertmanager** with incident management platforms like PagerDuty enhances operational responsiveness by ensuring that critical incidents are promptly addressed and resolved.

- **Understanding and addressing common ingestion issues**, such as consumer lag, broker failures, and message loss, is essential for maintaining robust and reliable data pipelines.

- **Exploring advanced anomaly detection methods** and implementing monitoring solutions like the Elastic Stack (ELK) can enhance the ability to manage and analyse complex data systems.

Thank you

Do you have any questions, comments, or feedback?