FACULDADE ENGENHARIA

Departamento de Informática

Plataformas e Serviços X-Ops (16233)

DataOps

Today’s Goals

- Cover the basics of DataOps
- Introduce the key components of DataOps
- Discover DataOps tools and platforms
- Hands-on activity

What is DataOps?

- ✧ DataOps is a practice focused on optimizing data analytics and engineering pipelines.
- ✧ It combines principles of data management and DevOps.
- ✧ DataOps improves efficiency, quality, and collaboration in data workflows.
- ✧ It's essential for organizations handling large-scale data and analytics.

Key Principles of DataOps

- ✧ **Agile** approach to data management
  - ▪ Rapid iteration and response to changing data needs.
  - ▪ Timely insights and reliable data processes.
- ✧ Continuous integration and delivery (**CI/CD**) for data
  - ▪ CI/CD in DataOps automates data integration and delivery, ensuring consistent and reliable data workflows.
- ✧ Cross-functional **collaboration**
  - ▪ Bridges data engineering, analytics, and operations for seamless data flow.
- ✧ **Automation**, monitoring, and quality control
  - ▪ Automation and monitoring reduce manual intervention, ensuring data accuracy and pipeline health.

Data Version Control

- ✧ Data version control is crucial for reproducibility and auditing in DataOps workflows.
- ✧ Similar to code versioning, it allows tracking changes in data.
- ✧ Popular tools include:
    - ▪ Git (for code)
    - ▪ DVC (Data Version Control)
- ✧ These tools help maintain data lineage and facilitate rollbacks.

Pipeline Orchestration

- Pipeline orchestration coordinates data flows between tasks and processes.
- Tools like Apache Airflow, Prefect, and Luigi manage dependencies and automate complex workflows in data pipelines.

Automated Testing in DataOps

- Automated tests ensure data quality by checking for consistency, accuracy, and schema compliance.
- Testing helps to catch errors early and maintain reliability.

Data Testing Strategies in DataOps

- ✧ Common strategies:
  - Data validation
  - Schema checks
  - Statistical tests
- ✧ These tests safeguard data quality at each step of the pipeline.

Data Governance in DataOps

- ✧ Data governance ensures security, privacy, and compliance in data operations.
- ✧ As exemple, tha GDPR guide data access, usage, and retention policies, ensuring compliance and data integrity.

Monitoring in DataOps

- Monitoring helps detect and resolve issues in real-time, optimizing pipeline performance.
- Monitoring is essential for early error detection.
- Tools like Prometheus, Grafana, and the ELK stack offer insights into data pipeline health and performance metrics.

DataOps Workflow

- ✧ DataOps workflows are designed to manage the entire lifecycle of data, from ingestion to insights.
- ✧ This demonstration provides a step-by-step guide through each phase.

Step 1 - Data Ingestion

- ✧ Data ingestion is the process of collecting raw data from various sources into a centralized repository.
- ✧ DataOps uses automated pipelines to streamline ingestion from different sources.
- ✧ Common tools include Apache Kafka, Apache NiFi, and Talend. These tools support both batch and real-time data ingestion.

Step 2 - Data Transformation

- ✧ Data transformation converts raw data into a structured format suitable for analysis.
- ✧ This includes cleaning, normalization, and enrichment processes.
- ✧ ETL tools like DBT, Apache Spark, and Talend handle data transformation. DBT is particularly popular for SQL-based transformations.

Step 3 - Data Storage and Versioning

- ✧ DataOps relies on storage solutions that support data versioning for traceability and rollback.
- ✧ Data is stored in data warehouses or lakes like Amazon S3, BigQuery, or Snowflake.
- ✧ DVC (Data Version Control) and Delta Lake provide tools to manage and version datasets. These tools are essential for tracking data changes over time.

Step 4 - Data Validation and Testing

- Data validation ensures data accuracy, consistency, and quality.
- Testing frameworks verify that data transformations produce the expected outcomes.
- Great Expectations, Deequ, and custom scripts are used for data validation. These tools help ensure data quality and integrity at each stage.

Step 5 - Monitoring and Observability

- Real-time monitoring detects issues and ensures pipeline health.
- Observability tools track metrics, errors, and system performance.
- Prometheus, Grafana, and ELK stack are used to monitor and visualize data pipeline health. They provide alerts and dashboards to quickly identify issues.

Hands-on activity

- **Flash Fiction Story** (**FFS)** is a style of very short storytelling that typically focuses on a single moment or idea, often 200-300 words.
- In the context of the classroom, it encourages creativity, improves concept retention, and promotes critical thinking by requiring students to distill complex technical ideas into concise stories.

Key Elements of Flash Fiction Story

✧ **Creativity**:

- Pushes students to think creatively, requiring them to condense complex ideas into a concise, impactful story.
- This is especially helpful for technical subjects where creative problem-solving is key.

Key Elements of Flash Fiction Story

✧ **Concept Retention**:

- Writing a story forces students to engage with the material more deeply, helping them remember core concepts.
- For example, they might write a short story about a data pipeline failure during a product launch, which helps them think through the possible causes and solutions.

Key Elements of Flash Fiction Story

✧ **Critical Thinking**:

- To tell a story in a few sentences, students need to focus on the essential elements of a scenario.
- This can lead to a deeper understanding of the topic as they prioritize key aspects of DataOps, such as monitoring, data validation, or automation.

Key Elements of Flash Fiction Story

✧ **Engagement**:

- Short stories add an element of fun and narrative to technical material, which can boost student engagement.
- Reading their stories aloud or sharing them in small groups adds a collaborative, interactive dimension to the learning experience.

Key Elements of **Flash Fiction Story**

✧ **Application in Real-World Scenarios**:

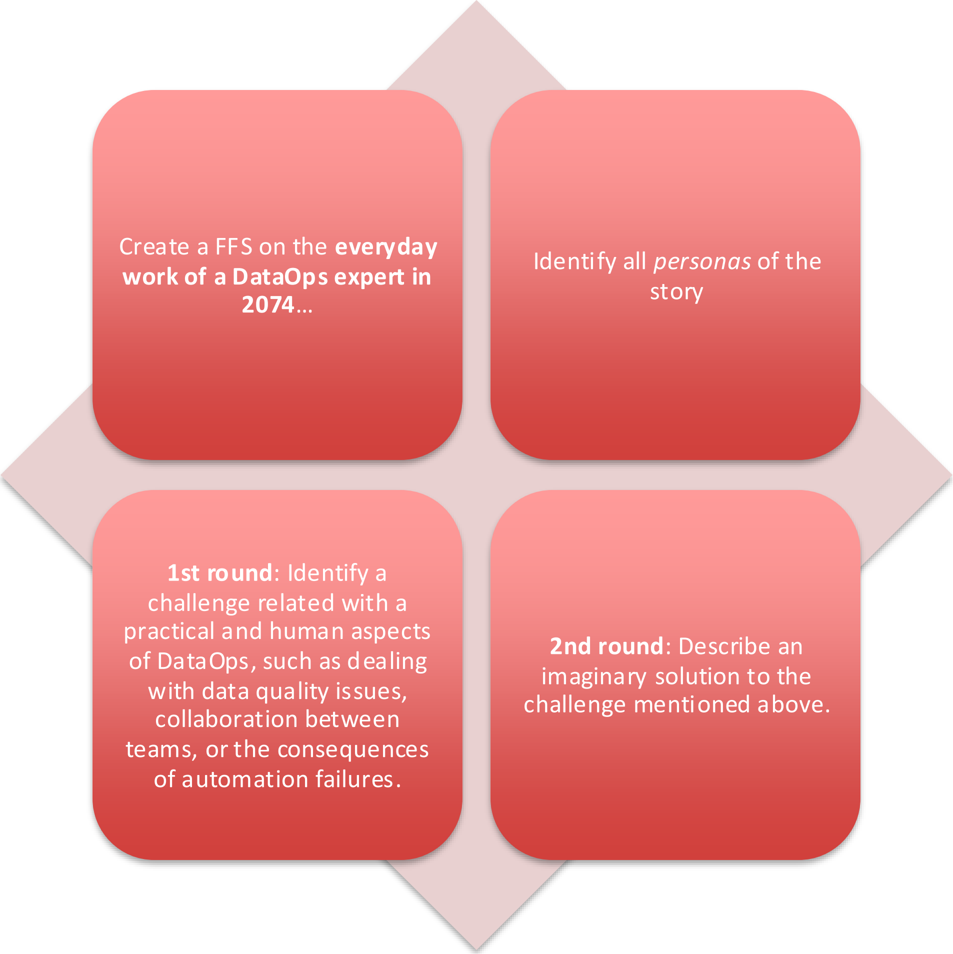
- A FFS can simulate real-world challenges by having students create narratives around hypothetical, yet realistic scenarios.
- For instance, into the DataOps context, they might write about troubleshooting a pipeline error right before a deadline, helping them consider the stress and quick thinking involved in operational roles.

Warm-up!

Create a FFS on the **everyday work of higher education student in 2054**…

- ✧ Write down his/ her name.
- ✧ **1st round**: Describe where and how he/she lives, and what he/she does on a normal workday.
- ✧ **2nd round**: Describe a positive workday of your main character.
- ✧ **3rd round**: Describe a negative workday of your main character.

Show-time!



Create a FFS on the **everyday work of a DataOps expert in 2074**…

Identify all *personas* of the story

1st round: Identify a challenge related with a practical and human aspects of DataOps, such as dealing with data quality issues, collaboration between teams, or the consequences of automation failures.

2nd round: Describe an imaginary solution to the challenge mentioned above.

Presentation Time!

3-minutes pitch

Think critically! (think about you would address the challenges presented in each story)

HOW
WHY
HOW
WHEN
WHERE
WHO
WHERE
WHERE
WHY
WHO
WHO
WHEN
WHAT
WHAT
WHY
WHAT
WHEN
HOW
WHAT
WHEN
WHO
WHERE
WHY
HOW