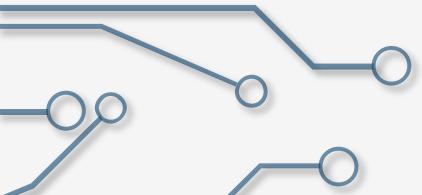
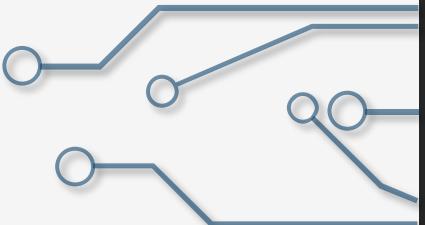
A decorative background pattern on the left side of the slide consists of a dense grid of blue lines forming a circuit board or network diagram, with small blue circles at various junction points.

DEVON JENNINGS  
PYDATA JOHANNESBURG  
MARCH 2024

# DEMYSTIFYING DATA INGESTION CHALLENGES WITH APACHE NIFI



WHO AM I?





# DEVON JENNINGS

INTERMEDIATE DATA ENGINEER  
AT CALYBRE

B.SC. QUANTITATIVE RISK  
MANAGEMENT  
(NORTH WEST UNIVERSITY)

I LOVE CROSSFIT !

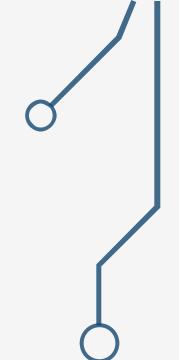


# AGENDA

- Introduction
  - Data Ingestion Challenges
  - Need For Speed in Data Ingestion
  - Influence of Open File Formats
  - Apache NiFi Demo
  - Question and Answers
- 
- 



# INTRODUCTION



# FACTS

- There are nearly as many pieces of digital information as there are stars in the universe.
  - Less than 0.5% of all data we create is ever used or analyzed.
  - According to PragmaticWorks, global businesses lose about 20-35% of their operating revenue from poor-quality data.
  - Poor data can cost the US government about \$3.1 trillion a year.
  - If you burned all the data created in 1 day into DVDs, you could stack them on top of each other to reach the moon twice.
- 
- 



# TRAFFIC JAM !



# DATA INGESTION CHALLENGES



## DATA QUALITY ISSUES

- Inconsistency and Incompleteness
- Data Validation



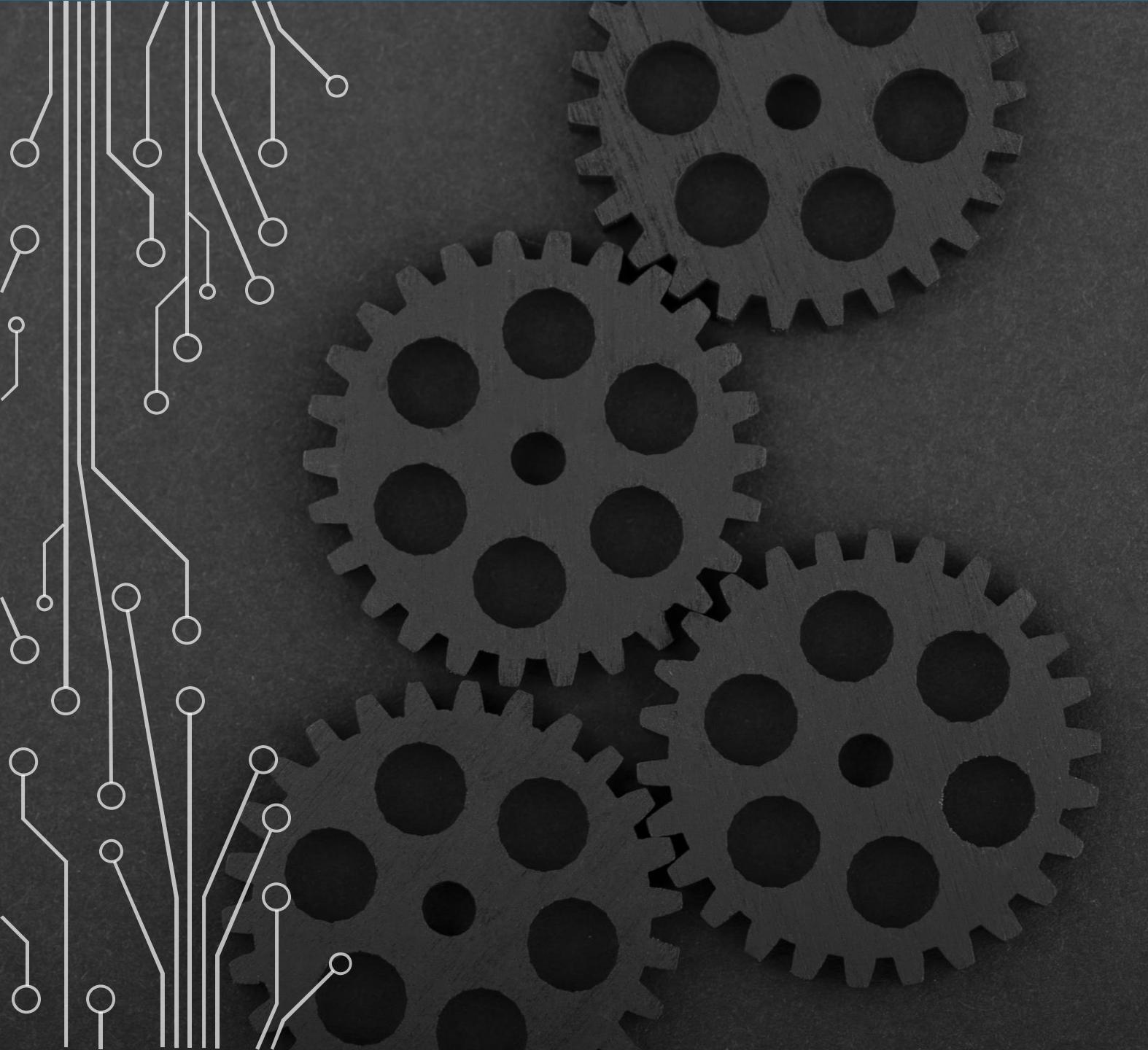
# DATA VOLUME AND VELOCITY

- Large Volumes of Data
- High Velocity



# DATA INTEGRATION ACROSS SOURCES

- Diverse Data Sources
- Schema Evolution



## **RELIABILITY AND FAULT TOLERANCE**

- Data Loss and  
Duplicates**
- Fault Tolerance**



# SECURITY AND COMPLIANCE

- Data Security
- Compliance Requirements



## ADDITIONAL CHALLENGES IN DATA INGESTION

- Scalability
- Data Governance
- Data Lineage and Traceability
- Data Transformation and Enrichment
- Data Synchronization
- Cost Management
- Metadata Management
- Handling Unstructured Data

# SUMMARY

- Data Quality Issues
- Data Volume and Volume
- Data Integration Across Sources
- Reliability and Fault Tolerance
- Security and Compliance



## NEED FOR SPEED IN DATA INGESTION



# INDUSTRIES AND USE CASES WHERE SPEED IS CRITICAL



FINANCE AND TRADING



HEALTHCARE



RETAIL

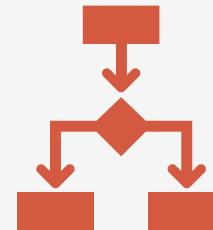


TELECOMMUNICATIONS

# IMPACT OF LATENCY IN REAL-TIME DATA SCENARIOS



**Operational Efficiency**



**Decision-Making**



**Customer Experience**

# COST IMPLICATIONS BEYOND INFRASTRUCTURE



MISSED  
OPPORTUNITIES



RESOURCE  
UTILIZATION



MAINTENANCE  
AND SUPPORT

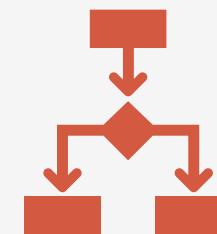


# INFLUENCE OF FILE FORMATS

# ROLE OF COLUMNAR STORAGE IN IMPROVING QUERY PERFORMANCE



**Data Organization**



**Query Performance**



**Parallel Processing**



# DATA ORGANIZATION



Row-Based Storage



Columnar Storage

# QUERY PERFORMANCE

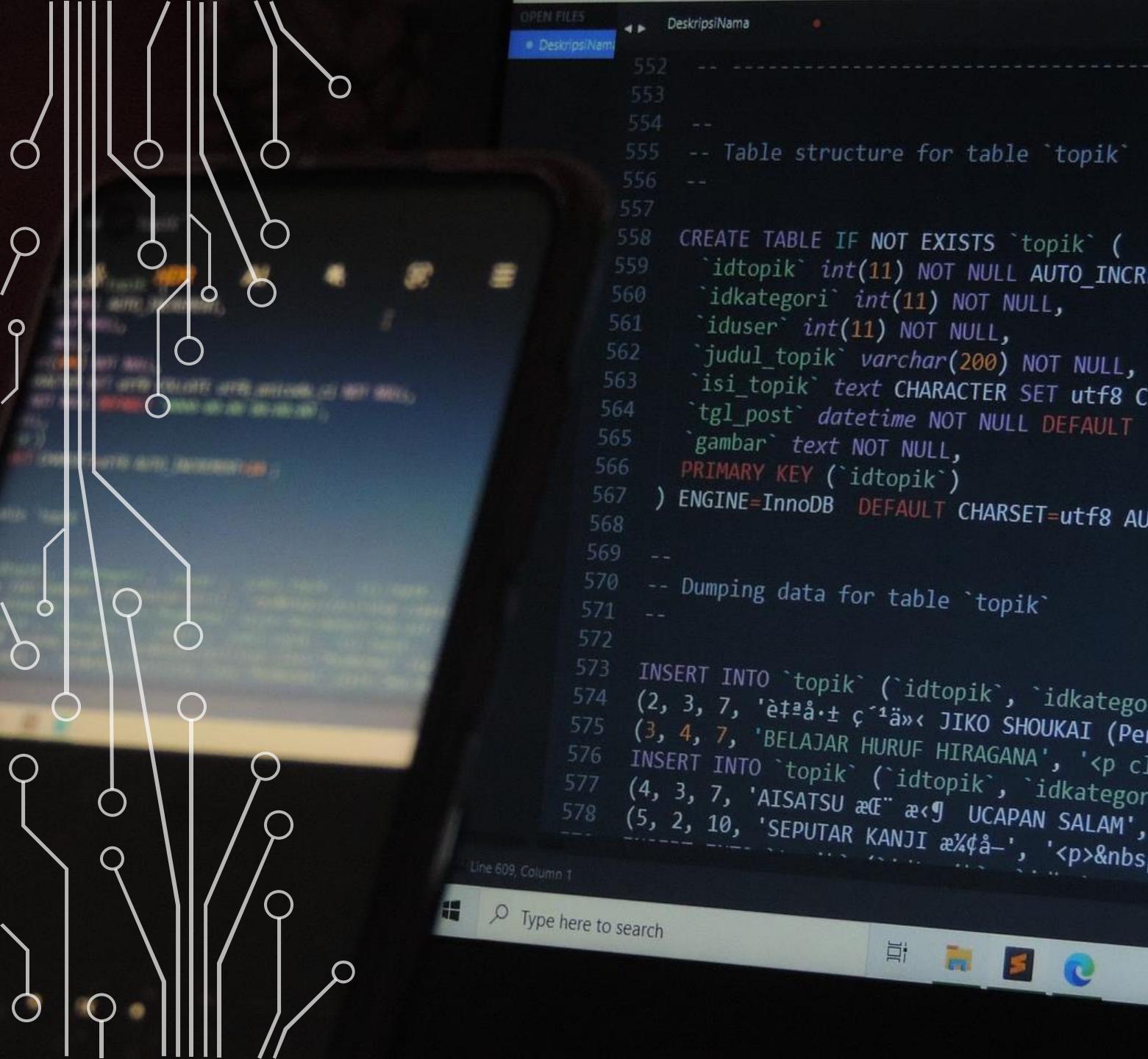


Selective Column Retrieval



Compression

```
OPEN FILES DesripsiNama
* DesripsiNama
552 --
553 --
554 --
555 -- Table structure for table `topik`
556 --
557
558 CREATE TABLE IF NOT EXISTS `topik` (
559   `idtopik` int(11) NOT NULL AUTO_INCREMENT,
560   `idkategori` int(11) NOT NULL,
561   `iduser` int(11) NOT NULL,
562   `judul_topik` varchar(200) NOT NULL,
563   `isi_topik` text CHARACTER SET utf8 COLLATE utf8_general_ci NOT NULL,
564   `tgl_post` datetime NOT NULL DEFAULT '0000-00-00 00:00:00',
565   `gambar` text NOT NULL,
566   PRIMARY KEY (`idtopik`)
567 ) ENGINE=InnoDB DEFAULT CHARSET=utf8 AUTO_INCREMENT=1000;
568 --
569 --
570 -- Dumping data for table `topik`
571 --
572
573 INSERT INTO `topik`(`idtopik`, `idkategori`, `iduser`, `judul_topik`, `isi_topik`, `tgl_post`, `gambar`)
574 (2, 3, 7, 'è‡·å·± ç·ä»« JIKO SHOUKAI (Perkenalan dan Belajar Huruf Hiragana)', '<p>classe
575 (3, 4, 7, 'BELAJAR HURUF HIRAGANA', '<p>classe
576 INSERT INTO `topik`(`idtopik`, `idkategori`, `iduser`, `judul_topik`, `isi_topik`, `tgl_post`, `gambar`)
577 (4, 3, 7, 'AISATSU æ€æ„æ„ UCAPAN SALAM', '<p>classe
578 (5, 2, 10, 'SEPUTAR KANJI æ%få–', '<p>&nbsp;
```





# PARALLEL PROCESSING



Parallelism



Vectorized  
Processing



## ADVANTAGES/DISADVANTAGES BETWEEN FILE FORMATS

- Avro
- Parquet

# AVRO



## Advantages

1. Schema Evolution
2. Compactness

## Disadvantages

1. No Compression
2. Slow Query Performance



# PARQUET



# Parquet

## Advantages

1. Columnar Storage
2. Compression

## Disadvantages

1. Schema Evolution
2. Write Performance

# APACHE NIFI DEMO



# INTRODUCTION TO APACHE NIFI



# WHY APACHE NIFI

Web-Based  
User Interface

Data Ingestion

Data  
Transformation

Flow Control

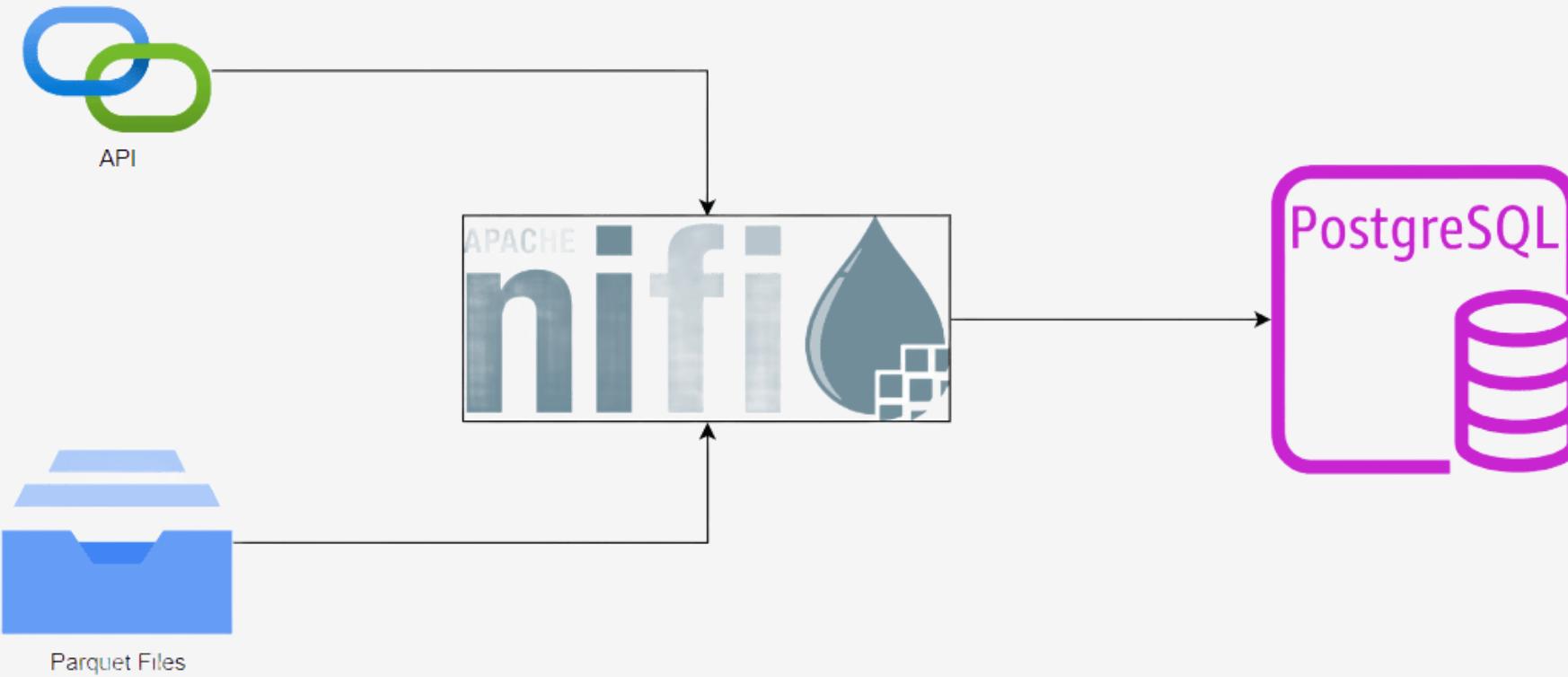
Security

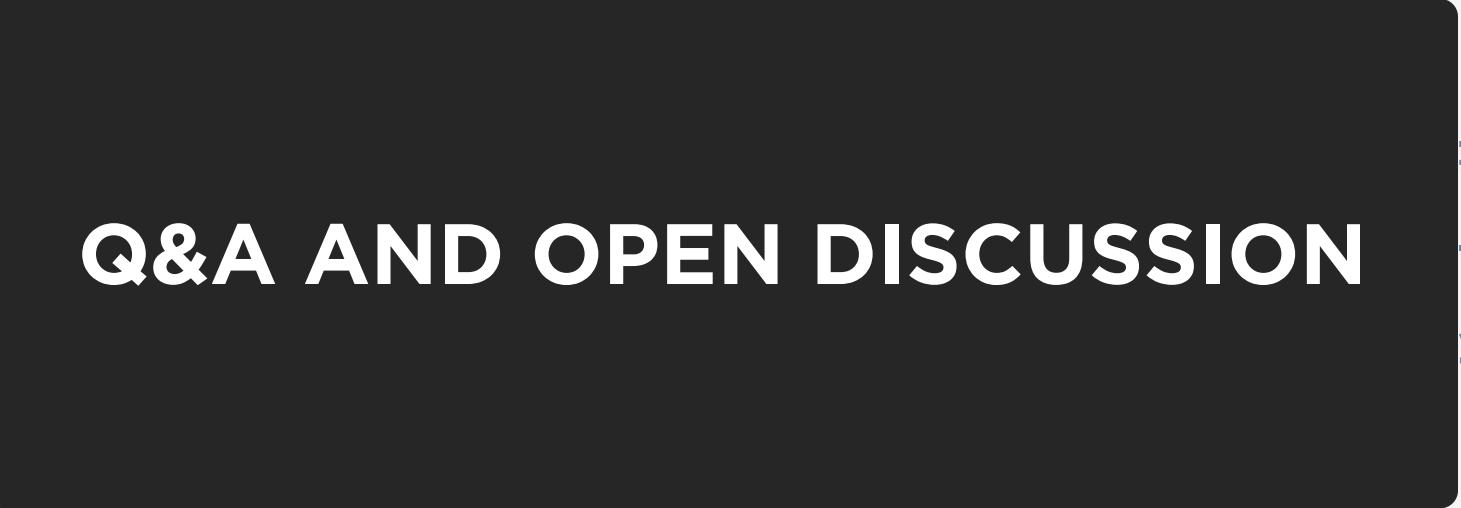
Extensibility

Data  
Provenance

Scalability

# ARCHITECTURE

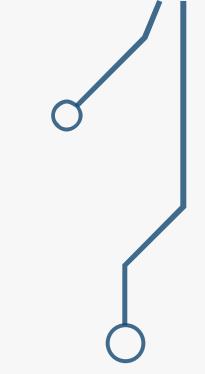




## **Q&A AND OPEN DISCUSSION**



**THANK YOU**



# SOURCES

- [https://www.pythian.com/blog/top-challenges-of-data-ingestion-pipelines-and-how-to-overcome-them-with-google-cloud]
  - [https://www.reply.com/net-reply-uk/en/content/challenges-in-data-gathering-and-the-ingestion-process]
    - [https://www.montecarlodata.com/blog-data-ingestion/]
  - [Devonjenn/DemistifyingDataIngestionChallengesWithApacheNifi](#)