PUBLIC TRANSPORT PIPELINE MONITORING

REAL-TIME INGESTION, STORAGE, AND MONITORING USING BIG DATA TOOLS

Guided by
Yiru Zhang

Presented by
Arthur AMUDA
Devkumar Parikshit GANDHI
LODHI Usama Pervez Khan

OBJECTIVE

OBJECTIVE

1

Objective 1

Build an end-to-end data pipeline for public transport data

2

Objective 2

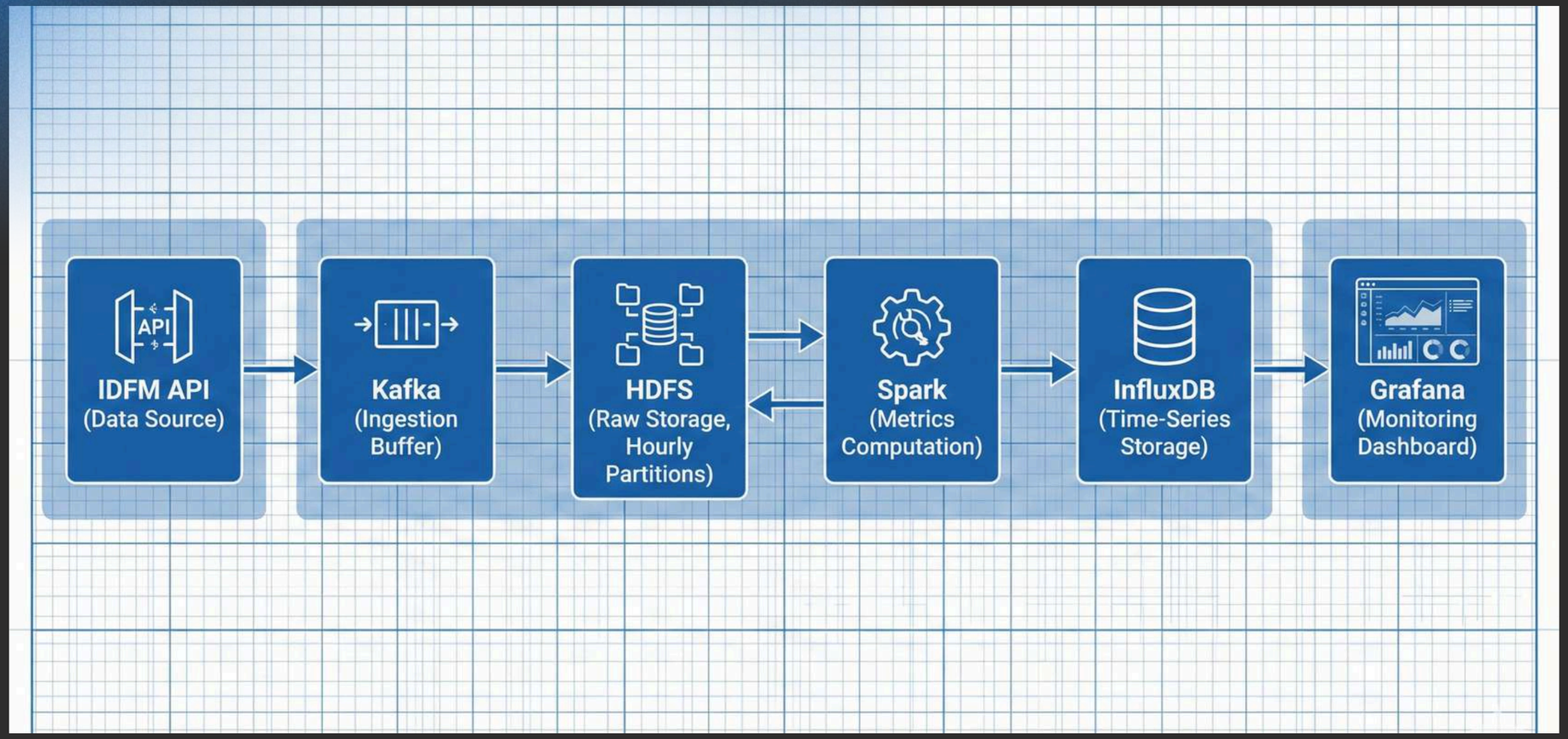
Monitor ingestion, storage, and processing metrics

3

Objective 3

Ensure scalability and observability

ARCHITECTURE OVERVIEW



DATA INGESTION & STORAGE

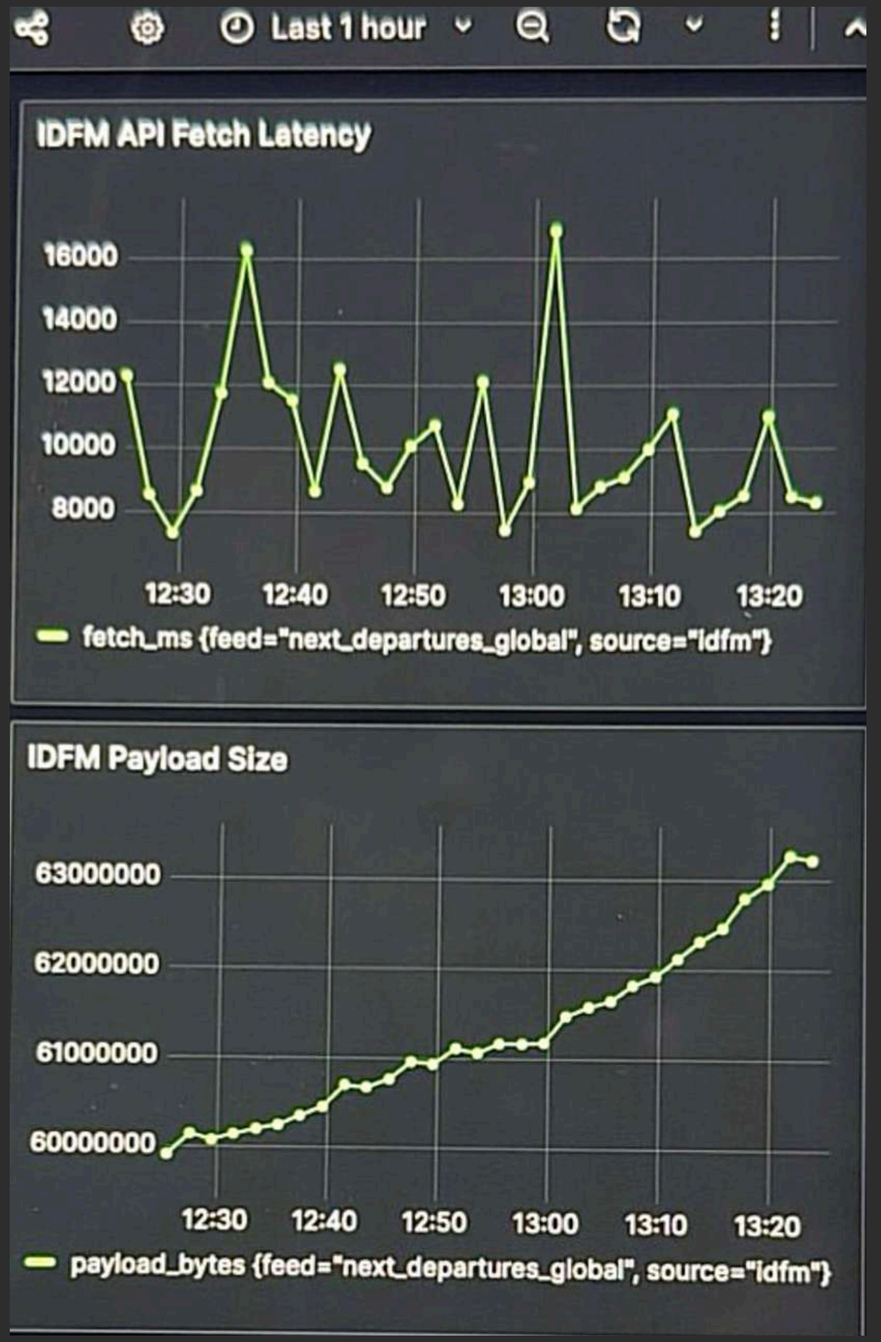
- JSON data fetched periodically from IDFM API
- Data written to HDFS
- Partitioned by:
- source
- date (dt)
- hour (hr)
- Enables scalable batch analytics

- Spark batch jobs compute hourly metrics:
  - Number of files ingested
  - Total bytes written to HDFS
  - Results exported to InfluxDB

ANALYTICS LAYER (SPARK)

MONITORING & DASHBOARD

- Centralized monitoring via Grafana
- Time-series visualization
- Metrics tagged by date, hour, and source



KEY METRICS OBSERVED

KEY METRICS OBSERVED

- Hourly file count
- Hourly total bytes written
- Storage growth trends over time

- Some metrics intentionally not activated:
  - API latency
  - Kafka offsets
  - Reason:
  - Stability
  - Time constraints
  - Focus on core KPIs

LIMITATIONS

FUTURE IMPROVEMENTS

- Enable latency tracking
- Add streaming analytics
- Alerting & anomaly detection
- Data quality checks

CONCLUSION

- Fully functional end-to-end pipeline
- Production-oriented architecture
- Clear observability through monitoring

```
WARNING: YARN_CONF_DIR has been replaced by HADOOP_CONF_DIR. Using value of YARN_CONF
_DIR.
2026-01-18 22:18:31,771 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to
 ResourceManager at master/10.0.0.46:8032
Total Nodes:4
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Con
tainers
   worker0:37697                RUNNING      worker0:8042
      0
   worker1:42429                RUNNING      worker1:8042
      0
   worker3:36283                RUNNING      worker3:8042
      0
   worker2:43077                RUNNING      worker2:8042
      0
```

CONCLUSION

THANKS

Presented by

Arthur AMUDA
Devkumar Parikshit GANDHI
LODHI Usama Pervez Khan