PROJETO FINAL
TECNOLOGIAS DE BIG DATA E INTELIGÊNCIA ARTIFICIAL

GRUPO 8 | INDIGESTION

Nome dos Alunos:
Alan Batista
Manuella Paez
Mario José C M Prado
Wislom Diogo Almeida

Coordenador:
Prof. Fabio Jardim

Agenda

1. Contextualização do trabalho
2. Visão e objetivo do projeto
3. Documentação da solução
    i. Diagrama da arquitetura e descrição dos serviços
    ii. Detalhamento e configurações técnicas
4. Demonstração da solução e entregáveis

1. Contextualização do trabalho

O sistema de transporte público de São Paulo atende **milhões de pessoas** diariamente, e a **eficácia na gestão da frota de ônibus** é crucial para **assegurar uma boa qualidade do serviço público** prestado e, consequentemente, a satisfação da população com o tema.

O trabalho possui como objetivo a construção de uma **aplicação** que possibilite o **monitoramento em tempo quase real dos ônibus** em circulação no estado de São Paulo e que **ofereça métricas e KPIs importantes para tomada de decisão**.

2. Visão e objetivo do Projeto

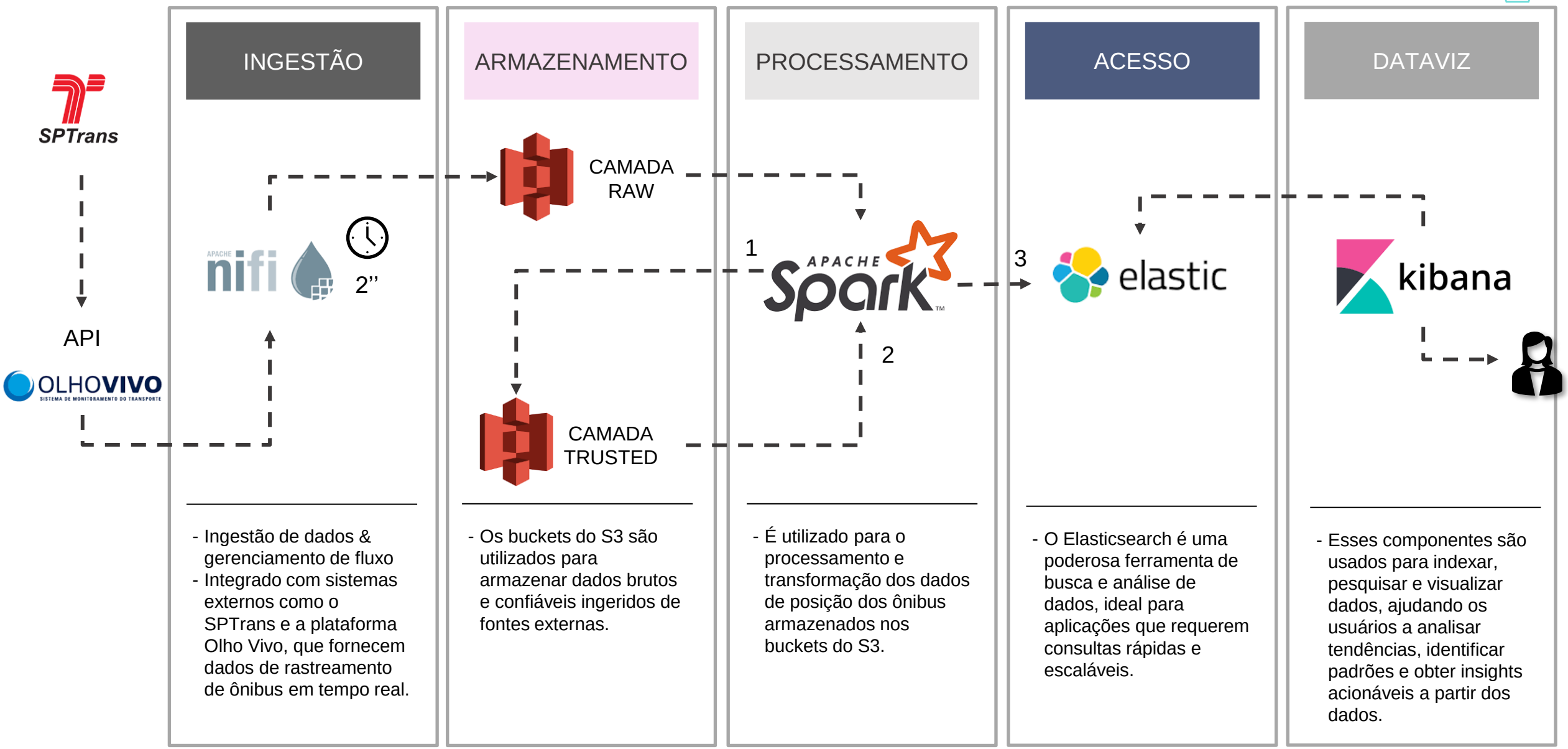
Solução de monitoramento em tempo real dos ônibus em circulação no cidade de São Paulo

Uso dos dados da API OLHO VIVO da SPTrans e dados GTFS

Integração com um data lake para armazenamento e enriquecimento de dados

Business intelligence para monitoramento em tempo quase real e análises para otimização do transporte público de São Paulo, Brasil.

3.i Solução | Arquitetura da solução e descrição dos serviços



- Ingestão de dados & gerenciamento de fluxo
- Integrado com sistemas externos como o SPTrans e a plataforma Olho Vivo, que fornecem dados de rastreamento de ônibus em tempo real.

- Os buckets do S3 são utilizados para armazenar dados brutos e confiáveis ingeridos de fontes externas.

- É utilizado para o processamento e transformação dos dados de posição dos ônibus armazenados nos buckets do S3.

- O Elasticsearch é uma poderosa ferramenta de busca e análise de dados, ideal para aplicações que requerem consultas rápidas e escaláveis.

- Esses componentes são usados para indexar, pesquisar e visualizar dados, ajudando os usuários a analisar tendências, identificar padrões e obter insights acionáveis a partir dos dados.

3.ii Solução | Detalhamento e configurações técnicas

INGESTÃO E ARMAZENAMENTO | PROCESSAMENTO | ACESSO E DATAVIZ

Credenciais de acesso:

NiFi:
- **URL**: http://localhost:49090
- **Porta padrão para acesso à interface web do NiFi :** 9090

MinIO:
Console: http://localhost:49001
Serviço de Armazenamento: http://localhost:49000
Portas expostas:
- 9000: Porta padrão para acesso ao serviço MinIO
- 9001: Porta do console de gerenciamento

Credenciais de Acesso:
Usuário: admin / **Senha**: minioadmin

GITHUB

1 nifi

Detalhamentos adicionais

Configuração:
Imagem Utilizada: apache/nifi:${NIFI_VERSION}
Container Name: nifi-otmzsp
Hostname: nifi-otmzsp

Ambiente:
NIFI_WEB_HTTP_PORT: 9090
NIFI_WEB_HTTPS_HOST: nifi
Timezone (TZ): America/Sao_Paulo

Volumes:
- *./volumes/nifi/util:/útil* | Diretório para utilitários.
- *./volumes/nifi/util/jar:/util/jar* | Diretório para arquivos JAR.
- *./volumes/nifi/conf:/opt/nifi/nifi-current/conf* | Diretório de configuração do NiFi.

Comando:
sh -c "ln -snf /usr/share/zoneinfo/$$TZ /etc/localtime && echo $$TZ > /etc/timezone“ | Configuração do timezone.

Recursos:
Limite de Memória: 2 GB para garantir desempenho adequado.

2

Configuração:
Imagem Utilizada: minio/minio:${MINIO_VERSION}
Container Name: minio-otmzsp

Estrutura de Pastas:
- Raw: Diretório para armazenamento de dados brutos.
- Trusted: Diretório para armazenamento de dados confiáveis.

Volumes: Mapeamento do diretório local *./volumes/minio/data* para */data* no container, garantindo persistência de dados.

Health Check:
- **Comando**: ["CMD", "mc", "ready", "local"]
- **Intervalo**: 5 segundos
- **Retries**: 5 tentativas
- **Timeout**: 5 segundos

3.ii Solução | Detalhamento e configurações técnicas

INGESTÃO E ARMAZENAMENTO | **PROCESSAMENTO** | ACESSO E DATAVIZ

Credenciais de acesso:

Apache SPARK:
- **URL**: spark://spark-master-otmzsp:7077

Portas expostas - Master:
- 7077: Porta do master para gerenciar os workers.
- 8080: Interface web do master.
- **Portas expostas - Worker:** 8081: Interface web do worker.

JUPYTER:
Portas expostas:
- 8888: Porta padrão para acessar a interface web do Jupyter Notebook.
- 4040 a 4043: Portas usadas para monitoramento do Spark (UI do Spark).

GITHUB

Detalhamentos adicionais

3 Apache Spark

Configuração:
Imagem Utilizada: apache/spark:${SPARK_VERSION}
Container Name:
- Master: *spark-master-otmzsp*
- Worker: *spark-worker-otmzsp*

Modos de Operação:
1. **Master**:

Variáveis de Ambiente:
SPARK_MODE: master
SPARK_MASTER_HOST: spark-master-otmzsp
TZ: America/Sao_Paulo
Limite de Memória: 2 GB.
Comando: /opt/spark/bin/spark-class org.apache.spark.deploy.master.Master

2. **Worker**:

Variáveis de Ambiente:
SPARK_MODE: worker
SPARK_MASTER_URL: spark://spark-master-otmzsp:7077
SPARK_WORKER_MEMORY: 1g
TZ: America/Sao_Paulo
Limite de Memória: 1 GB.
Comando: /opt/spark/bin/spark-class org.apache.spark.deploy.worker.Worker spark://spark-master-otmzsp:7077

4 jupyter

Configuração:
Imagem Utilizada: *jupyter/pyspark-notebook:latest*
Container Name: *jupyter-otmzsp*
Ambiente:
JUPYTER_TOKEN: "" | Desabilita o token de segurança para acesso.

Volumes:
../notebooks:/home/jovyan/work | Diretório para armazenar notebooks, permitindo persistência de dados e fácil acesso.

Comando:
start-notebook.sh --NotebookApp.token='' --NotebookApp.password=” | Inicia o servidor Jupyter sem token ou senha, facilitando o acesso.

3.ii Solução | Detalhamento e configurações técnicas

INGESTÃO E ARMAZENAMENTO | PROCESSAMENTO | ACESSO E DATAVIZ

Credenciais de acesso:

Elastic Search:
Portas expostas:
- 9200: Porta padrão para acesso à API REST do Elasticsearch.
- 9300: Porta para comunicação entre nós (cluster).

Kibana:
Portas expostas:
- 5601: Porta padrão para acessar a interface web do Kibana.

GITHUB

5 elastic

Detalhamentos adicionais

Configuração:
Imagem Utilizada: elasticsearch:7.17.20
Container Name: elasticsearch-otmzsp
Hostname: elasticsearch-otmzsp

Ambiente:
discovery.type: single-node | configuração para executar em modo de nó único)
.ES_JAVA_OPTS: "-Xms2g -Xmx2g" | configurações de memória do Java
.xpack.security.enabled: "false" | desabilita a segurança para simplificar a configuração.

Volumes:
./volumes/elasticsearch/esdata:/usr/share/elasticsearch/data | Mapeamento para persistência de dados.

Health Check:
Comando: curl -sS --fail http://elasticsearch-otmzsp:9200/_cluster/health?wait_for_status=yellow&timeout=0s
Intervalo: 1 segundo
Retries: 3 tentativas
Start Period: 20 segundos
Timeout: 5 segundos

6 kibana

Configuração:
Imagem Utilizada: kibana:7.17.20
Container Name: kibana-otmzsp
Hostname: kibana-otmzsp

Ambiente:
ELASTICSEARCH_HOSTS: "http://elasticsearch-otmzsp:9200" | configuração para conectar ao Elasticsearch.

Volumes:
./volumes/kibana/data:/usr/share/kibana/data | Mapeamento para persistência de dados, garantindo que as configurações e dashboards sejam mantidos.

Dependências:
Elasticsearch: O Kibana depende do Elasticsearch, sendo necessário garantir que o serviço do Elasticsearch esteja saudável antes do início do Kibana.

PROJETO FINAL
DATA ENGINEERING

GRUPO 8 | INDIGESTION

Nome dos Alunos:
Alan Batista
Manuella Paez
Mario José C M Prado
Wislom Diogo Almeida

Coordenador:
Prof. Fabio Jardim