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

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

Agenda

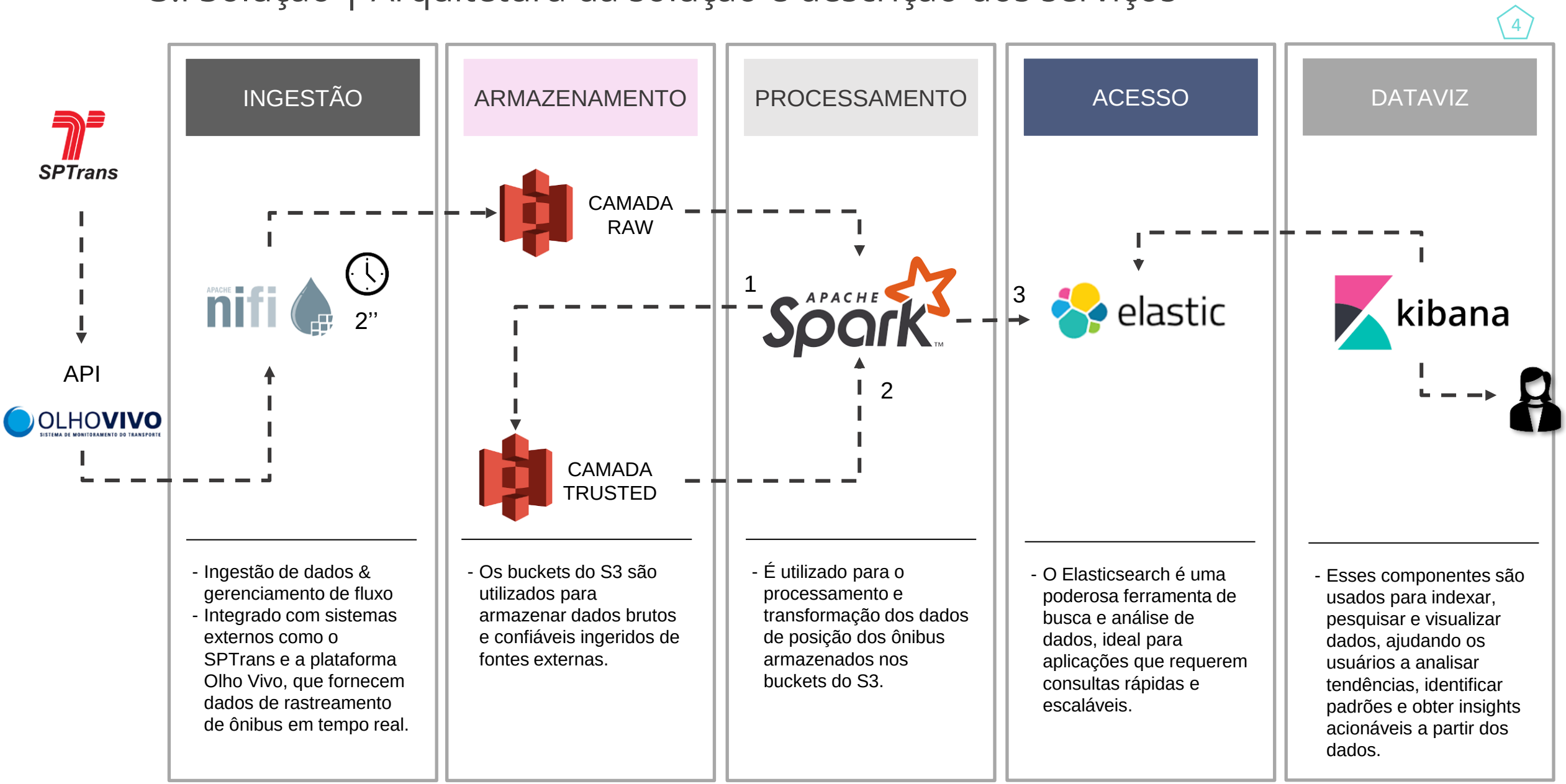
1. Contextualização do trabalho
2. Visão e objetivo do projeto
3. Documentação da solução
    i. Diagrama da arquitetura e descrição dos serviços
    ii. Detalhamento e configurações técnicas
4. Demonstração da solução e entregáveis

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**.

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



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 montados para persistência de dados: /minio_data/raw e /minio_data/trusted .

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**: http://localhost:8080 (UI do Spark Master)

Portas expostas - Master:
- 7077: Porta do master para gerenciar os workers.
- **URL**: http://localhost:8081 (UI do Spark Worker)
- **Portas expostas - Worker:** 8081: Interface web do worker.

JUPYTER:
- **URL**: http://localhost:8888

Portas expostas:
- 8888: Porta padrão para acessar a interface web do Jupyter Notebook.
- 4040 a 4043: Portas usadas para monitoramento 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:
• URL: http://localhost:49200 (API)

Kibana:
• **URL: http://localhost:45601**

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 single node
.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.

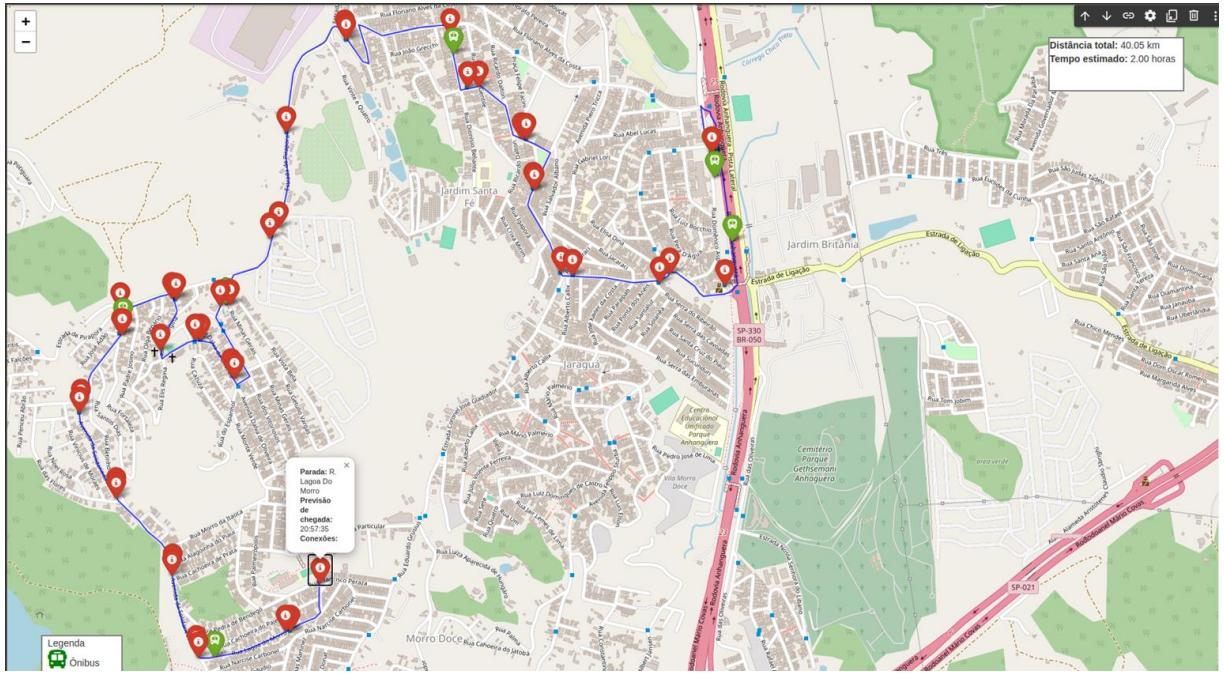
4. Entregáveis | Explorando a melhor solução

Exploramos opções, mas concluímos pela tecnologia que possibilitou a entrega na qualidade e tempo combinados

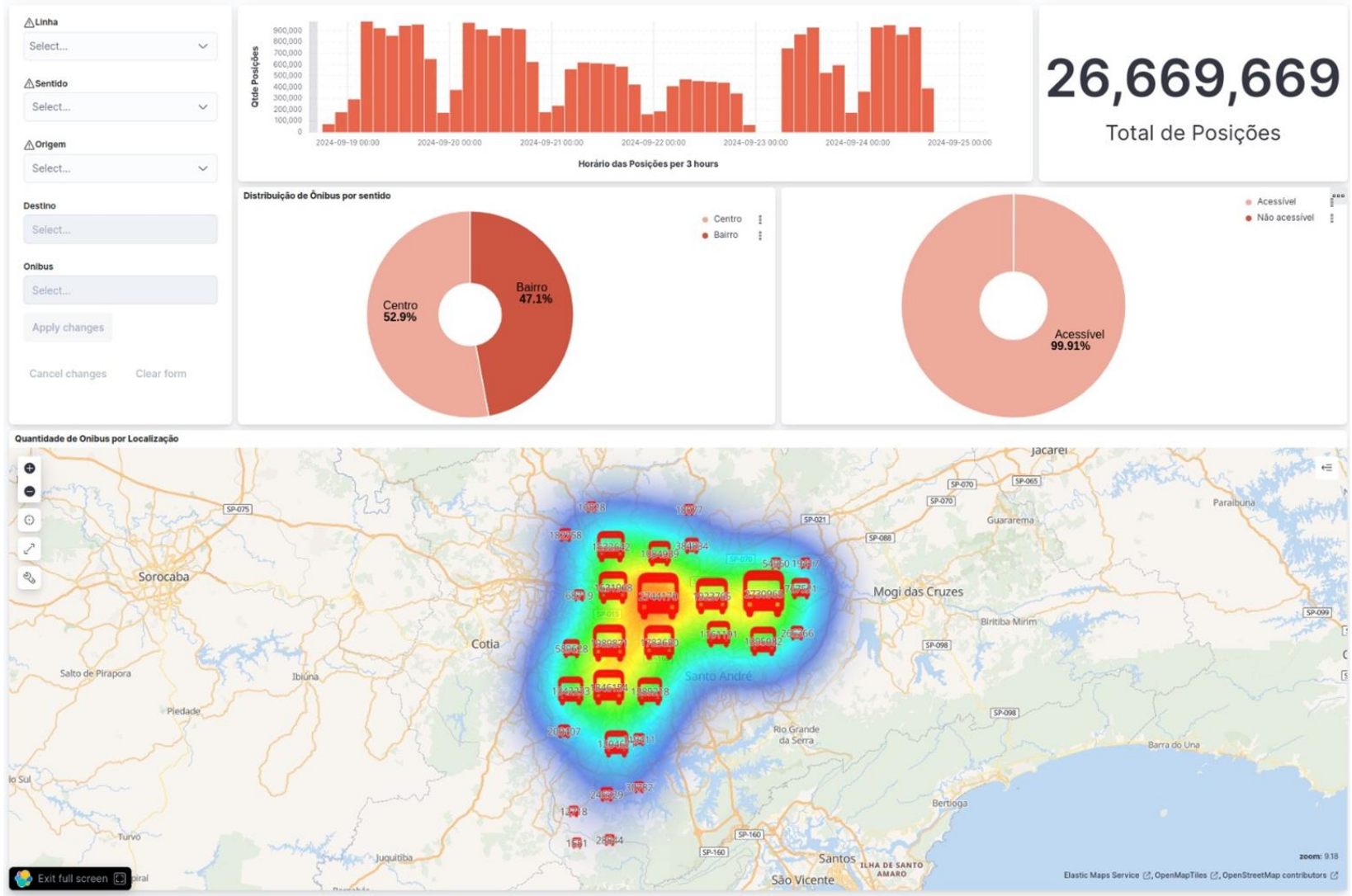
- Análises prévias incluíram a inspeção dos dados, detecção de padrões, e transformações visando a otimização dos fluxos de dados e a preparação para futuras integrações.

- No entanto, devido às limitações de tempo, não foi possível implementar todas as melhorias e conclusões dessas análises no pipeline atual.

Demonstração dos resultados que obtivemos, representando o percurso de uma linha de ônibus com estimativas de chegada para cada parada.



4. Entregáveis | Componentes e insights a partir do Dashboard final



1. Gráfico de Barras (Posições por Período de 3 horas) - mostra a variação do número de posições capturadas, refletindo a intensidade do tráfego de ônibus em diferentes períodos ao longo de vários dias.

2. Total de Posições - representa o volume total de registros de posições dos ônibus, acumulados durante o período de monitoramento.

3. Distribuição de Ônibus por Sentido (Centro/Bairro) - No exemplo, a proporção está próxima de 53,17% em direção ao Centro e 46,83% em direção ao Bairro.

4. Acessibilidade - 99,91% dos ônibus são Acessíveis, enquanto uma pequena fração (0,09%) não o é. Isso reflete a inclusão de veículos com recursos para pessoas com mobilidade reduzida.

5. Mapa de Calor (Heatmap) da Quantidade de Ônibus por Localização - facilita a visualização de regiões de maior tráfego, permitindo uma análise espacial da distribuição de ônibus na região metropolitana de São Paulo e cidades próximas, como Santo André, Cotia, e Mogi das Cruzes.

4. Entregáveis | Próximos passos

1. Ampliar a análise contemplando outros indicadores, principalmente relacionados à eficiência:
- Densidade de veículos por linha
- Quantidade de paradas por linhas
- Comprimento das linhas
- Frequência média de ônibus por linha
- Quatidade de linhas que passam pelo mesmo ponto
- Pontualidade
- Média de carros em circulação vs. média geral

1. Cruzar com outras fontes de dados externas para por exemplo, conseguir quantificar informações sobre custo do transporte.

1. Avaliar o investimento na feature de linha tracejadas no mapa da Kibana para avaliar melhor padrões de movimentação e cobertura.

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