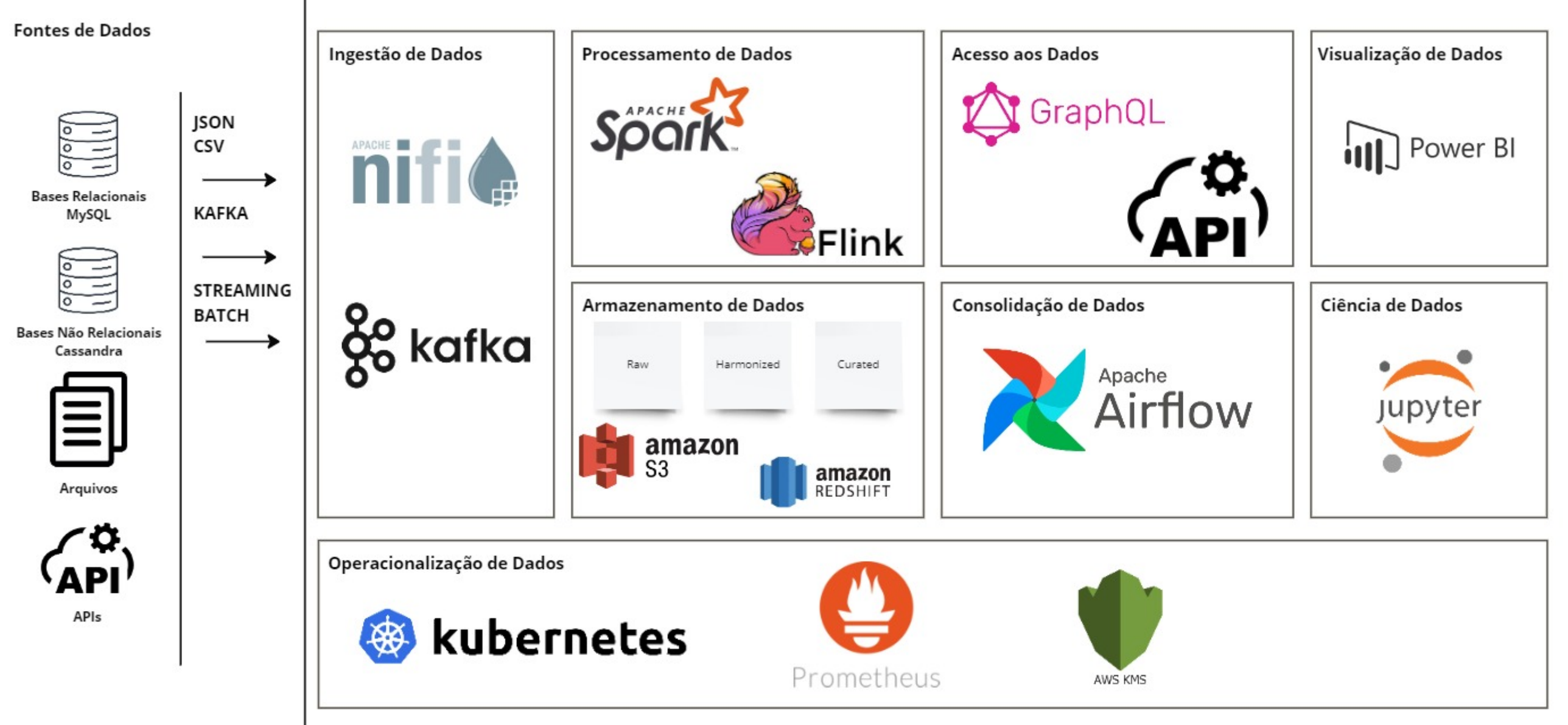
Fontes de Dados
Bases Relacionais MySQL
Bases Não Relacionais Cassandra
Arquivos
APIs
JSON
CSV
KAFKA
STREAMING
BATCH
Ingestão de Dados
APACHE nifi
kafka
Processamento de Dados
APACHE Spark
Flink
Acesso aos Dados
GraphQL
API
Visualização de Dados
Power BI
Armazenamento de Dados
Raw
Harmonized
Curated
amazon S3
amazon REDSHIFT
Consolidação de Dados
Apache Airflow
Ciência de Dados
jupyter
Operacionalização de Dados
kubernetes
Prometheus
AWS KMS

JUSTIFICATIVA DAS ESCOLHAS TECNOLÓGICAS

Fonte de Dados

Bases Relacionais (MySQL)
Justificativa: MySQL foi escolhido por ser um sistema de banco de dados relacional amplamente reconhecido pela sua eficiência e confiabilidade. É especialmente adequado para dados estruturados e operações transacionais, oferecendo suporte robusto a SQL para consultas detalhadas e assegurando a integridade dos dados, o que é crucial para transações financeiras.

Bases Não Relacionais (Cassandra)
Justificativa: A escolha do Cassandra é ideal para armazenar grandes volumes de dados distribuídos. Como um banco de dados NoSQL, ele proporciona alta disponibilidade e escalabilidade, sendo excelente para gerenciar dados em tempo real e volumes elevados de escrita, como logs e informações de usuários.

Arquivos e APIs (JSON, CSV, KAFKA, STREAMING, BATCH)
Justificativa: O uso de arquivos em formatos como JSON e CSV facilita a integração com diferentes sistemas e o armazenamento de dados semi-estruturados. Kafka é uma ferramenta chave para o processamento de dados em tempo real e ingestão eficiente, enquanto o processamento em batch é adequado para lidar com grandes volumes de dados em períodos específicos.

Ingestão de Dados

Apache NiFi: É eficaz na movimentação e gerenciamento de fluxos de dados entre diferentes sistemas. Com sua interface gráfica intuitiva, ele simplifica a configuração e supervisão dos processos de ingestão.

Apache Kafka: Destaca-se na ingestão e processamento de dados em tempo real, suportando altas taxas de transferência e baixa latência, o que é crucial para aplicações que necessitam de dados atualizados continuamente.

Ciência de Dados

Jupyther Notebook: Oferece um ambiente interativo perfeito para explorar dados, desenvolver e testar modelos de machine learning e criar visualizações. Ele facilita a experimentação e documentação do trabalho de análise de dados.

Processamento de Dados

Apache Spark: Boa opção para o processamento de dados em larga escala, capaz de lidar com tanto operações em batch quanto em tempo real, o que o torna ideal para análises complexas e grandes volumes de dados.

Apache Flink: É ideal para processamento contínuo de dados, oferecendo baixa latência e alta capacidade de throughput, essencial para sistemas que requerem análise de dados em tempo real.

Dados Processados
- RAW: Dados brutos coletados sem processamento inicial.
- Harmonized: Dados ajustados e integrados para análise.
- Curated: Dados refinados e preparados para relatórios e visualizações.

Acesso aos Dados

GraphQL: Permite consultas flexíveis e eficientes, garantindo que as APIs retornem apenas os dados necessários e evitando problemas comuns.

API RESTful: Oferecem uma forma estruturada e eficiente de integrar e acessar dados de diferentes sistemas e aplicações.

Armazenamento de Dados

Amazon S3: É uma solução escalável e durável para armazenar grandes volumes de dados não estruturados e backups.

Amazon Redshift: É um data warehouse poderoso que facilita a análise de grandes quantidades de dados com rapidez, integrando-se bem com ferramentas de visualização para fornecer insights detalhado.

Consolidação de Dados

Apache Airflow: É uma ferramenta eficiente para orquestrar e monitorar workflows de dados, facilitando o agendamento e gerenciamento de processos complexos de ETL.

Visualização de Dados

Power BI: Ideal para criar dashboards interativos e relatórios detalhados, permitindo a análise e compartilhamento de insights de forma visual e intuitiva.

Operacionalização de Dados

Kubernetes: Simplifica o gerenciamento e a escalabilidade de aplicações containerizadas, garantindo que os recursos sejam utilizados de maneira eficiente e que a aplicação esteja sempre disponível.

Prometheus: É fundamental para monitorar o desempenho e a saúde da infraestrutura, fornecendo métricas e alertas para garantir que os sistemas estejam operando conforme o esperado.

AWS KMS: Oferece uma camada adicional de segurança para criptografar dados em repouso, garantindo a proteção e conformidade com requisitos regulatórios.