ECOSSISTEMA HADOOP E ARQUITETURA DE *DATA LAKE*, EM UMA ABORDAGEM TECNOLÓGICA E PARA NEGÓCIOS COM FOCO EM *ANALYTICS*.

Um estudo de caso da empresa fictícia Foodhood.

MBA Executivo em Business Analytics e Big Data
MRJ02021 - TBABD-8 - Banco de Dados Distribuídos – Profª Fernanda Bruno

****

Rafael Henrique Rocha de Souza
A56660250

INTRODUÇÃO

Big data é o termo usado para descrever, de maneira mais simplista, grandes e complexos conjuntos de dados, que são difíceis de armazenar e processar, mas são um trampolim para a criatividade e tomada de decisão quando boas práticas analíticas são utilizadas. No entanto, apesar do avanço das tecnologias, seu gerenciamento e armazenamento ainda são tarefas desafiadoras. O conceito de *Data Lake*, alterou a forma em que tratamos esses volumes expressivos de dados e vem ajudando as empresas a manter dados brutos e díspares em seu formato nativo. Nesse repositório de big data, ou arquitetura de dados unificada, independente de como queira chamar, o que fica evidente, é sua facilidade de consolidação e integração, absorvendo dados não estruturados e criando uma acessível, segura, escalável e distribuída estrutura, pronta para consultas e análises.

O objetivo desse projeto é trazer uma discussão, com foco gerencial, na montagem de uma arquitetura de *Data Lake* em um ecossistema da plataforma Hadoop, apresentar suas camadas de tecnologias e o papel de cada uma.

Para enriquecer o estudo, será aplicado o caso da empresa fictícia **Foodhood** e seus desafios de mercado, uma *foodtech* nacional, idealizada exclusivamente para este trabalho. Ela teria como principal objetivo competir com as atuais líderes do mercado de delivery de comida pela internet: iFood e UberEats. Com seus imaginários quatro anos de história e um crescimento exponencial, ela hoje já geraria quase cinco bilhões de registros e alguns *terabytes* de dados por mês, o que sem dúvida se refletiria em um problema de negócio bastante complexo e perfeito para ser atacado com uma gestão mais eficaz de seus dados, e inicialmente, o uso de técnicas de business intelligence.

A FOODHOOD, SEUS DADOS E SEU PROBLEMA DE NEGÓCIO

A empresa tinha como principal desafio lidar com a heterogeneidade dos seus dados, ou seja, dados estruturados, semi e não estruturados. Outro problema era a necessidade de armazenamento de dados gerados por inúmeras fontes, que com seu sistema tradicional já não era mais viável, por uma razão óbvia, seus dados estavam aumentando em uma velocidade tremenda. Outra questão seria a velocidade de acesso e processamento, levando em consideração que a capacidade de seu disco rígido vinha aumentando, mas seu poder de processamento e capacidade de acesso não vinha crescendo na mesma taxa.

A IMPLEMENTAÇÃO DO ECOSSISTEMA HADOOP E SUAS FERRAMENTAS

A escolha da plataforma Hadoop pelos executivos da Foodhood, se fez por ela possibilitar o processamento desses grandes volumes de dados entre diversos computadores usando modelos simples de programação. Suas principais características são:

- Escalabilidade horizontal;
- Ser um *commodity hardware* mais barato, mesmo que com gerenciamento complexo;
- A vantagem de ser sem *schema* ou estrutura rígida;
- Ser durável, o que garante persistência e o poder de lidar com os dados mesmo com possíveis falhas de máquina;
- Ser capaz de fazer balanceamento dos dados.

O vasto volume de dados gerados pela Foodhood, vem surgindo como uma oportunidade. Seus executivos sabem que podem obter preciosos benefícios utilizando ferramentas de *Big Data Analytics and Business Intelligence*, podendo através delas

descobrir tendências de mercado, preferências de clientes, padrões ocultos, correlações desconhecidas, melhorar a eficiência operacional, trazer vantagens competitivas sobre as concorrentes e outras inúmeras informações relevantes ao benefício do negócio.

O Hadoop é uma estrutura que permitiu a Foodhood armazenar os dados em um ambiente distribuído, para que pudesse processá-lo paralelamente com seus dois principais componentes, o HDFS (*Hadoop Distributed File System*), que permite armazenar dados de vários formatos em um *cluster*; e o YARN que cuida do gerenciamento de recursos no Hadoop, permitindo o processamento paralelo dos dados, armazenados no HDFS.

O HDFS passa uma sensação de abstração, semelhante à virtualização, mesmo que logicamente apareça como uma unidade única de armazenamento de *big data*, ele na verdade armazena esses dados em vários nós, de uma forma distribuída, e segue a arquitetura de mestre-escravo, como ilustrado na figura abaixo:

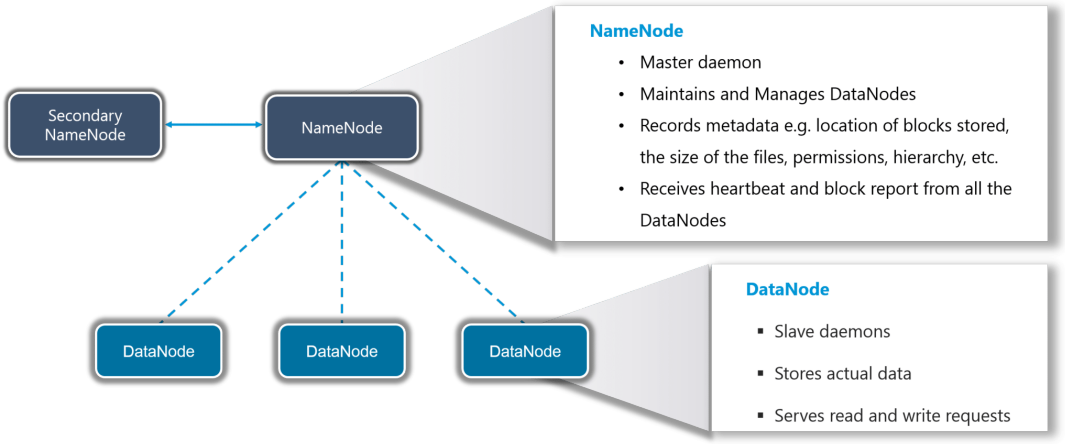

Figura 1. Como funciona o Hadoop – HDFS

Nesse bloco decidi listar uma série de recursos e benefícios que o HDFS trouxe para a empresa:

- **CUSTO:** Levando em consideração que o ecossistema foi implantado em um hardware comum, ele trouxe custo inicial baixo, visto que foi utilizado um commodity hardware e não foi necessário gastar muito dinheiro para dimensionar o cluster, adicionar mais nós ao HDFS de certa forma será barato para a Foodhood quando necessário;

- **ALTO RENDIMENTO:** O rendimento tem relação com a quantidade de trabalho realizado em uma unidade de tempo, rapidez nos acessos e alto desempenho no sistema com seu processamento dos dados em paralelo;

- **INTEGRIDADE DOS DADOS E CONFIABILIDADE:** O sistema verifica com constância a integridade dos dados da empresa, se encontrar alguma falha reporta ao nó responsável sobre isso e em seguida cria uma réplica adicional e logo exclui as cópias corrompidas. O fato do armazenamento ser distribuído traz confiabilidade. Cada *NameNode* gerencia os metadados e os *DataNodes* são responsáveis por armazenar os dados, eles são replicados e várias cópias são criadas, por padrão a replicação é de 3, e bastante consistente, isso torna o HDFS muito confiável. Vale mencionar que quando armazenado, por exemplo 1GB de arquivo, ele ocupará na verdade 3GB de espaço;

- **CONCEITO DE *DATA LOCALITY*:** Esse conceito se relaciona com a possibilidade de mover estrategicamente a unidade de processamento para os dados, ao invés do contrário, que era o caso antes da implantação do ambiente para a Foodhood, o que reduzia o desempenho da rede de forma perceptível. Agora trazemos a parte computacional para os nós, onde os dados residem;

- **VARIEDADE E VOLUME DE DADOS:** Se tornou possível armazenar qualquer tipo de dados no HDFS, volumosos no caso, em *terabytes* ou *petabytes*, sejam eles estruturados, não estruturados ou semi.

O YARN executa todas as suas atividades de processamento alocando recursos e programando tarefas, utilizando seus dois principais componentes: o *ResourceManager* e o *NodeManager*. O primeiro funciona como um nó mestre, que recebe as solicitações de processamento e as passa para os *NodeManagers* correspondentes, para que o processamento real ocorra, eles são instalados e se responsabilizam pela execução das tarefas de cada *DataNode*.

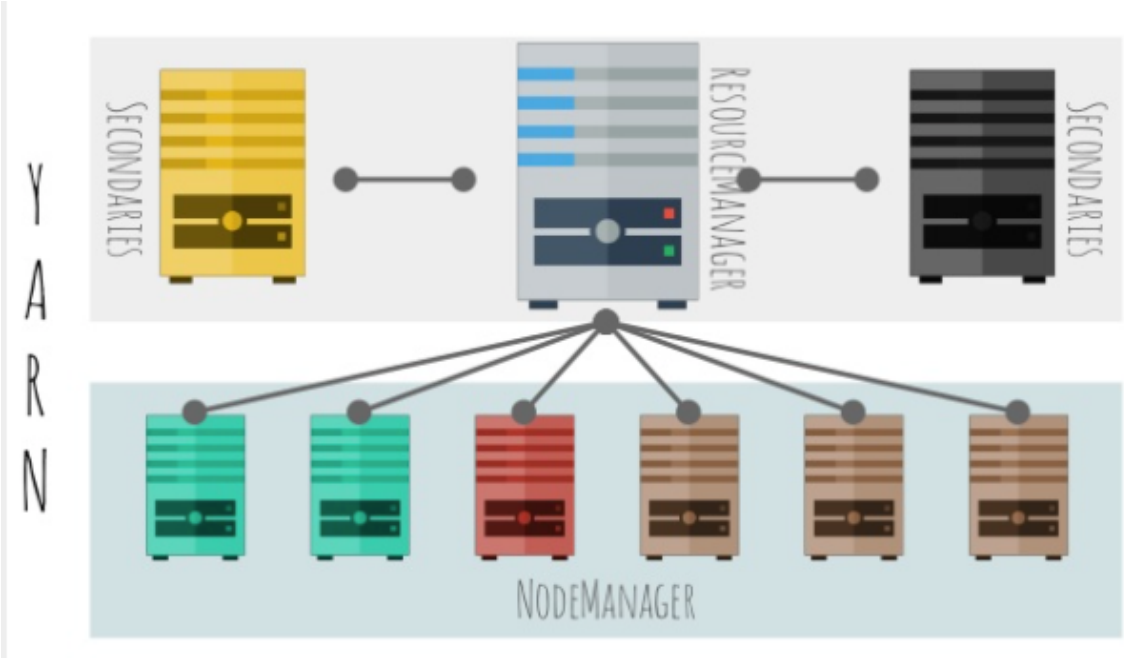

Figura 2. Funcionamento do YARN.

O YARN permite diferentes métodos de processamento de dados armazenados no HDFS, como de gráfico, interativo, de fluxo e de lote. Ele abre o Hadoop para outros tipos de aplicativos distribuídos além do MapReduce e permitiu que os usuários na Foodhood executassem operações conforme demanda, usando uma variedade de ferramentas que iremos detalhar daqui a pouco, como Spark para processamento em tempo real, Hive para SQL, HBase para NoSQL e outros. Além do gerenciamento de recursos ele também realiza agendamento de tarefas. A arquitetura Apache Hadoop YARN, consiste em Resource Manager, Node Manager, Aplication Master e Container,

como componentes principais. Esse último, por exemplo, apresenta o pacote de recursos de cada nó, incluindo, RAM, CPU, rede, HDD, entre outros.

O MapReduce, que é um aplicativo que fica em uma camada acima do YARN, é um modelo de programação muito versátil, que permite processar os dados em um cluster, com seus *reducers*, que são responsáveis por agregar os dados, e seus *mappers*, que tem a capacidade de transformar seus dados de maneira muito eficiente. *Mappers* e *reducers* juntos podem ser usados para resolver problemas complexos. Atualmente, com o Hadoop 2.0, foi introduzido um novo modelo de processamento que será discutido mais a frente.

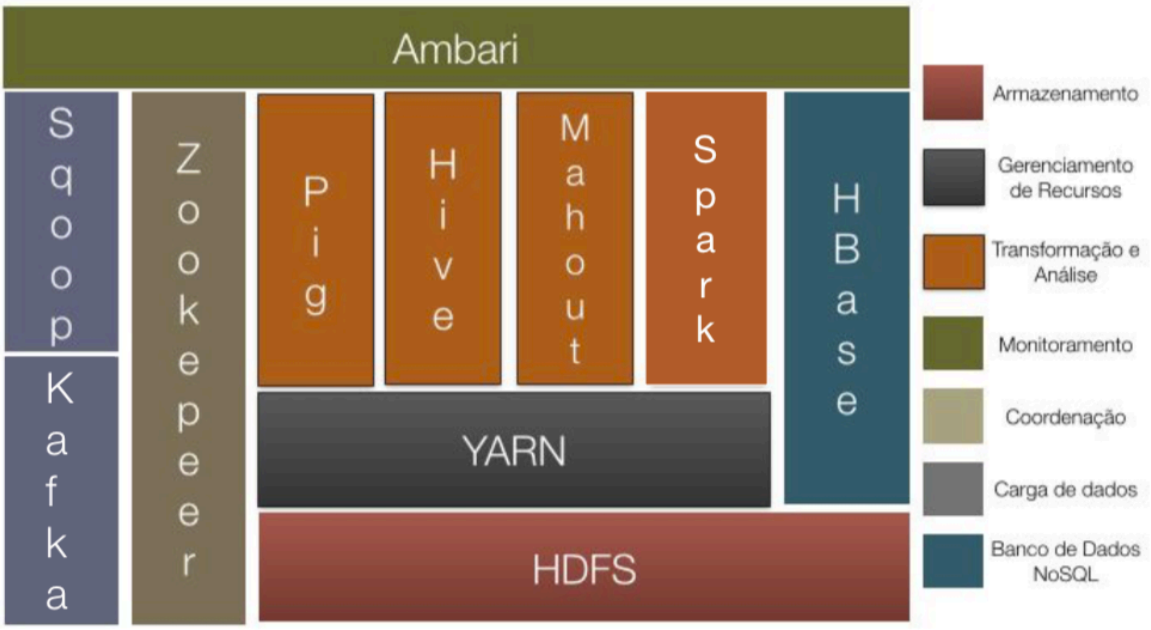

Figura 3. Ecossistema mínimo - Hadoop

Levando em consideração o ecossistema mínimo para a montagem de uma estrutura na plataforma Hadoop, como ilustrado na figura acima, serão listadas abaixo cada ferramenta utilizada e seus papéis:

- **AMBARI:** Ele oferece uma visão do estado real de seu cluster, em termos de aplicativos, e é basicamente uma ferramenta de administração e armazenamento responsável por rastrear aplicativos e manter seus status disponíveis. Ele permite que você visualize o que é executado em seu cluster, quais sistemas e quantos recursos estão sendo usados. É uma ferramenta de gerenciamento que se responsabiliza pelos motores e funcionamentos dos clusters;

- **SQOOP E KAFKA:** O SQOOP é uma ferramenta usada para transferir dados entre servidores de banco de dados relacional e é usado para importar dados de bancos relacionais como Oracle para Hadoop HDFS, MySQL e etc, e para exportar de HDFS para bancos de dados relacionais. O Kafka também é um software de processamento de dados de streaming e é usado para construir *pipelines* de dados em tempo real e aplicativos de streaming, reduzindo a complexidade. É horizontalmente escalável, tolerante a falhas e visa fornecer uma plataforma unificada de baixa latência para lidar com feeds de dados em tempo real. Comunicações e mensagens assíncronas podem ser estabelecidas com sua ajuda, o que garante uma comunicação confiável;

- **ZOOKEEPER:** É basicamente uma tecnologia que coordena tudo no cluster e pode ser usada para rastrear os nós ativos e inativos. Tem uma maneira muito confiável de controlar os estados compartilhados em seu cluster que diferentes aplicativos podem usar, e é responsável por manter um desempenho consistente de um cluster;

- **PIG:** Ferramenta de BI. Uma API de programação de alto nível que se baseia no MapReduce e permite escrever scripts simples. Disponibiliza respostas

complexas e executa funções de MapReduce sem realmente a necessidade de se escrever códigos do Java e sim semelhantes ao do SQL;

- **HIVE:** Também se baseia no MapReduce e é uma ferramenta tradicional de BI. Ele traz uma maneira fácil de se fazer consultas SQL e fazer com que os dados distribuídos em seu sistema de arquivos em algum lugar, pareçam um banco de dados SQL. Sua linguagem é conhecida com Hive SQL e possibilita consultas SQL nos dados armazenados em seu cluster Hadoop, embora não seja um banco de dados relacional subjacente;

- **MAHOUT:** Permite armazenar e processar *big data* em um ambiente distribuído em clusters usando modelos de programação simples. É uma estrutura de mineração de dados que normalmente é executada em conjunto com a infra do Hadoop em seu segundo plano para gerenciar grandes volumes de dados. Pode ser usado para criar algoritmos de *machine learning* escalonáveis e possibilita implementação de técnicas como recomendação, classificação e *clustering*;

- **SPARK:** Sem dúvidas a tecnologia mais interessante do ecossistema. Também possibilita execução de consultas nos dados e é o principal mecanismo de processamento de dados em tempo real, traz análises rápidas e mais fáceis de usar do que o MapReduce. É extremamente rápido e uma tecnologia muito poderosa, pois usa o processamentos de dados na memória e pode lidar com aprendizado de máquina em um cluster inteiro de informações, lidar com dados de streaming e etc. Como o PIG, HIVE E O MAHOUT é também uma ferramenta de transformação e análise;

- **HBASE:** Ele é definido lateralmente e é uma maneira de expor os dados em seu cluster à plataforma transacional. É chamado de banco de dados NoSQL, ou seja,

um armazenamento de dados colunar, muito rápido e destinado a grandes volumes de transações. Expõe dados armazenados em seu cluster que podem ser transformados de alguma forma pelo Spark ou MapReduce e fornece uma maneira muita rápida de expor esses resultados a outros sistemas.

Abaixo uma simples ilustração, criada pelo autor desse projeto, apresentando como foram incluídas as camadas de tecnologias consideradas relevantes no caso Foodhood, sempre com a visão de complementação ao BI:

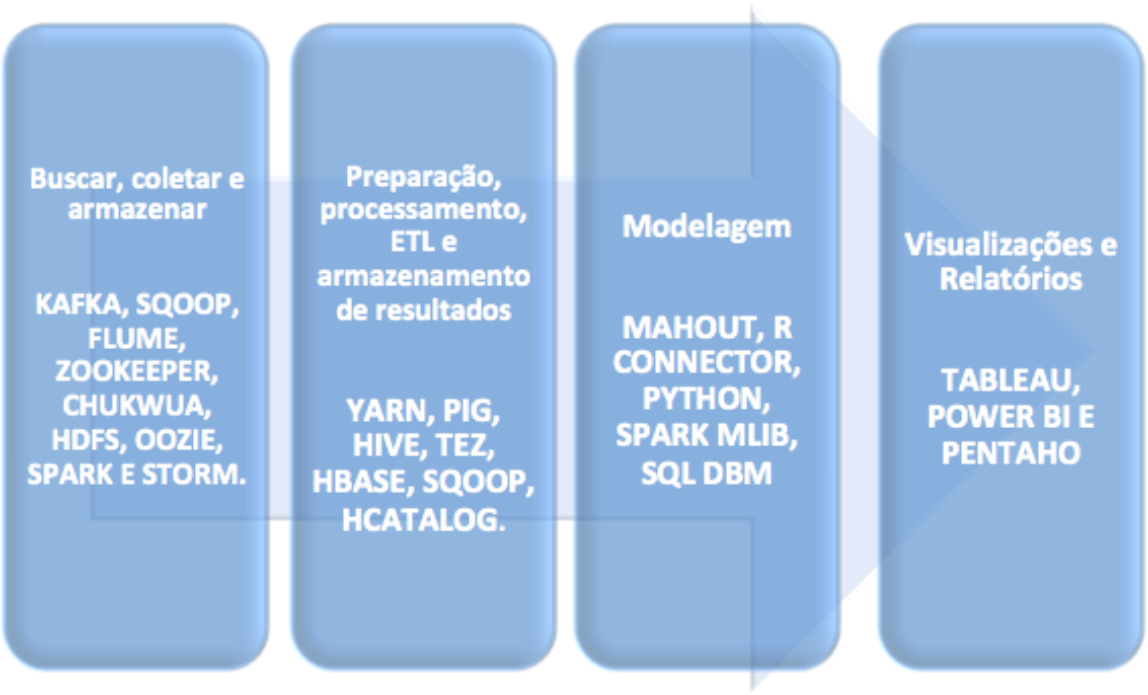

Figura 4 . Ecossistema Hadoop e fluxo BI para a Foodhood – Suas camadas e tecnologias.

O *DATA LAKE* DA FOODHOOD

O processo de construção do *Data Lake* da Foodhood, para que se tornasse robusto, foi recomendado um movimento gradual, com uso das ferramentas certas e com a plataforma cuidadosamente planejada. Foram tomados os seguintes cuidados:

- **Manipulação e ingestão dos dados escalonáveis:** Essa primeira fase envolveu a criação de uma estrutura, onde se aprendeu a adquirir dados em escala. Nessa fase, as análises são simples, consistindo em simples transformações;

- **Aumento da capacidade analítica dos usuários da Foodhood:** Nessa segunda fase concentrou-se em aprimorar a análise e interpretação dos dados e implantação de ferramentas e estruturas específicas que dessem suporte as demandas do negócio;

- **Integração entre as plataformas:** Nesse momento uma onda de democratização tomou conta da Foodhood, quando a estrutura já permitia análises, consultas e a inteligência já flutuava livremente pela empresa. Este é o estágio onde ocorreu uma sinergia completa e contínua com o Data Lake e a utilização da arquitetura;

- **Adoção de ponta a ponta e aquisição de maturidade:** Aqui ocorreu o estágio final e superior de maturidade, onde houve um real aumento da capacidade da empresa e da governança dos dados, segurança, auditoria, gerenciamento dos metadados e do ciclo de vida das informações geradas e insights.

RESULTADO E GERAÇÃO DE VALOR

Agora com seu Data Lake em operação, a Foodhood pode trazer dados de várias fontes, independentemente do volume, e se tornou capaz de executar análises descritivas e preditivas e dessa forma dar suporte aos tomadores de decisão. Seu fluxo de dados é contínuo no repositório, o nível de segurança aumentou significativamente e hoje tem escalabilidade dos dados.

A orquestração dos dados traz maior agilidade no manuseio e consultas, e consequentemente melhor integração com ferramentas de BI. Nossa empresa fictícia adquiriu com esse investimento, prontidão para lidar com desafios futuros relacionados a dados e análises, e no geral essa solução trouxe a cultura analítica para o core da empresa e sem dúvida capacidade para competir no mercado com seus grandes concorrentes e tomar decisões mais assertivas.

BIBLIOGRAFIA

SHARDA, Ramesh; DURSUN, Delen; EFRAIN, Turban. **Business Intelligence e Análise de Dados Para Gestão de Negócio**. 4. ed. Porto Alegre: Bookman. 2019. 584 p.

PROVOST, Foster; FAWCETT, Tom. **Data Science Para Negócios.** 1. ed. Rio de Janeiro: Alta Books. 2016. 384 p.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados.** 6. ed. São Paulo: Pearson. 2005. 502 p.

GOLERIK, Alex. **The Enterprise Big Data Lake.** 1. ed. Califórnia: O'Reilly. 2019.

APACHE HADOOP, 2020. Disponível em: https://hadoop.apache.org/