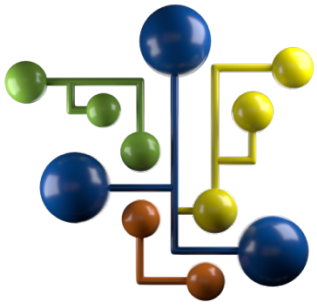


Data Science Academy

www.datascienceacademy.com.br

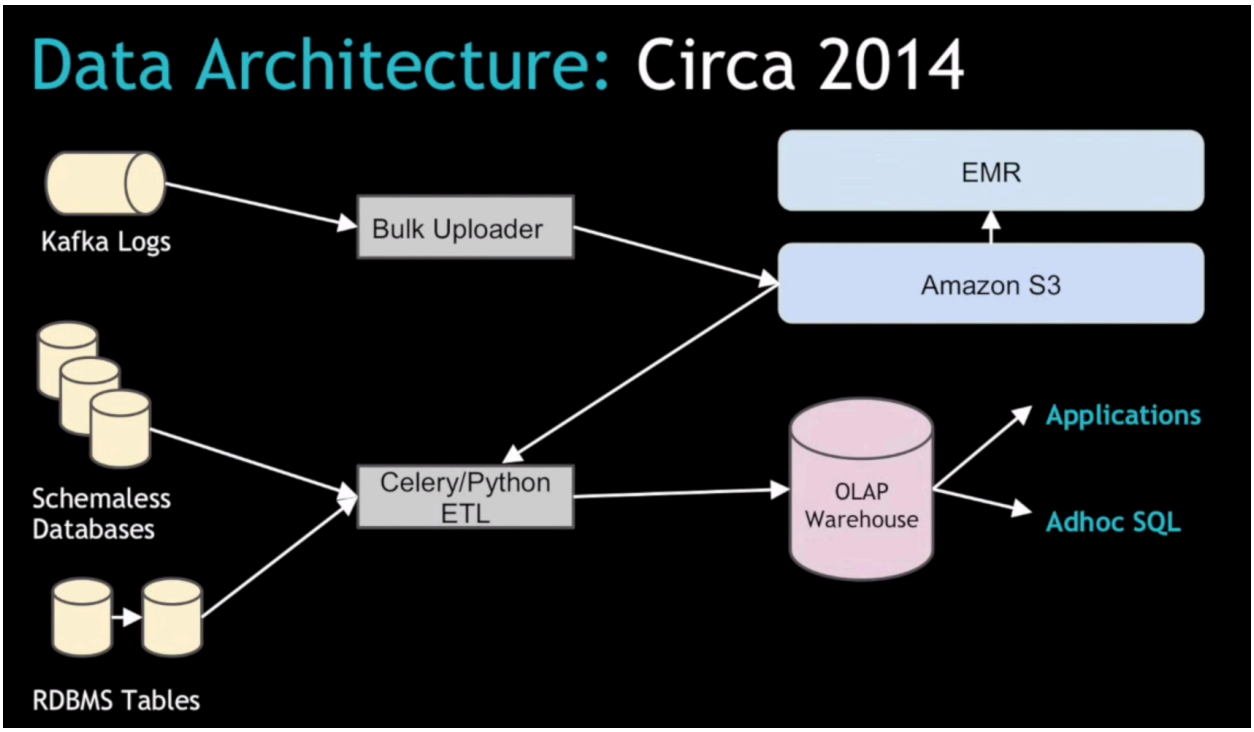
Big Data Real-Time Analytics com Python e Spark

Como o Uber utiliza Spark e Hadoop

Se você alguma vez utilizou o Uber, sabe como o serviço é simples. Você pressiona um botão, um carro aparece, você vai até o seu destino, aperta outro botão para pagar o motorista e pronto! Obviamente, existe uma engenhosa infraestrutura permitindo que o processo seja tão simples. E o Uber utiliza Spark e Hadoop para fazer isso acontecer.

O Uber está na interseção entre o mundo virtual e o mundo físico. Com mais de 160 mil motoristas cadastrados, 8 milhões de usuários e cerca de 2 bilhões de viagens, o Uber confia no processamento em tempo real, mais do que em qualquer outra tecnologia. Processamento de dados em tempo real é alma do negócio do Uber. Em uma recente mesa redonda com a equipe da Databricks, o "Head of Data" do Uber, Aaron Schildkrout declarou: "Nosso negócio é fundamentalmente um problema de dados". (Link com as estatísticas completas na seção de links úteis).

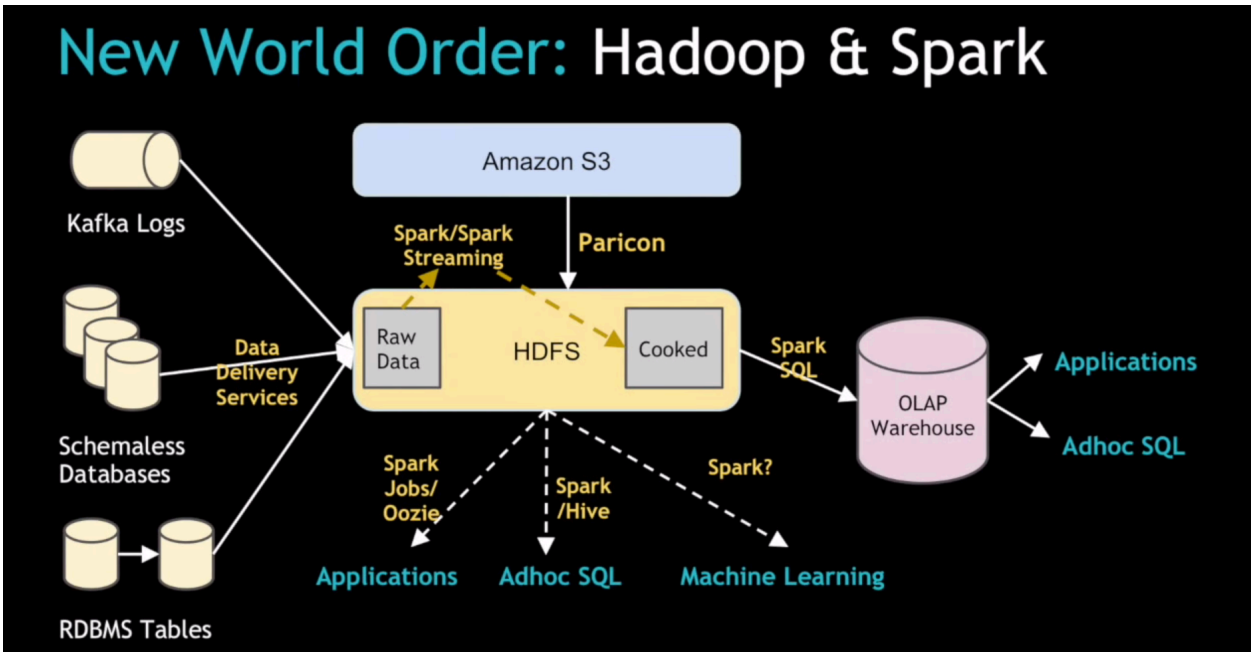
Os engenheiros do Uber recentemente descreveram os desafios na coleta e manipulação dos dados que a empresa enfrenta e como eles atendem a demanda cada vez maior de usuários. Inicialmente o Uber utilizava em sua infraestrutura uma ferramenta ETL baseada em Python (Celery), o Kafka para coleta de logs e o Amazon S3 para armazenamento em nuvem. Os dados eram processados pelo EMR (Amazon Elastic MapReduce). Essa solução funcionava bem, mas o Uber começou a ter problemas de escala (crescimento da infraestrutura) à medida que novas cidades eram atendidas pelo serviço, com a expansão do Uber pelo mundo. Os maiores problemas estavam no processamento em batch de grandes volumes de dados.



Arquitetura inicial do Uber

A solução encontrada pelos engenheiros foi a criação de uma nova infraestrutura baseada no Spark, que substituiu a ferramenta de ETL Celery/Python e o Hadoop HDFS para armazenamento dos dados. A ideia era coletar e armazenar os dados em seu estado bruto (sem qualquer tratamento inicial) e isso foi feito com o HDFS e então usar o Spark para processamento em larga escala. Somente no momento do processamento que os dados eram organizados e transformados para análise.

Assim, a nova infraestrutura ao invés de tentar agregar os dados das viagens de múltiplos data centers, em bancos relacionais, usava o Kafka para coletar o Stream de logs das mudanças nos dados a partir dos data centers locais e carregar tudo isso em um cluster Hadoop. O sistema então usa o Spark SQL para converter dados sem schemas (não estruturados) em dados com schemas no formato JSON e então aplicar análise de dados usando SQL e HQL (Hive SQL). Os engenheiros trabalharam ainda em um sistema de ingestão de dados usando o Spark Streaming, de modo a poder aplicar modelos de Machine Learning. Eles ainda estão trabalhando nesta última solução.



O Spark é agora a porta de entrada para coleta de dados de diversas fontes (Kafka, bancos de dados relacionais e dados não estruturados). O Spark usa o HDFS como sistema de armazenamento e com isso tem a possibilidade de utilizar ferramentas Hadoop, com o Hive por exemplo. O Spark SQL coleta os dados e envia para os Data Warehouses, enquanto o processamento e análise de dados em tempo real é feito com o Spark Streaming e modelos de Machine Learning com MLLib (solução ainda em desenvolvimento).

O que você estudou e viu ao longo dos capítulos sobre Spark, é o que está sendo usado em uma das empresas mais inovadoras do mundo!

Obrigado

Equipe Data Science Academy

Fonte:

Spark and Spark Streaming at Uber - Meetup
https://www.youtube.com/watch?v=zKbds9ZPjLE