



# DATA LAKE

DESIGN, PROJETO E INTEGRAÇÃO





# Data Lake

## Aquisição de Dados em Batch





# Data Lake - Aquisição de Dados em Batch

The diagram features a central vertical blue line. To the left, a green line branches out to a light green circle. To the right, a yellow line branches out to a light yellow circle. At the bottom, an orange circle is connected to the central line. Three blue rounded rectangles are positioned around the central line, each containing text about different data lake architectures.

Data Lake On-premises  
(Usando Máquinas Virtuais)


Data Lake na Nuvem  
(Usando AWS)

Serverless Data Lake  
(Usando AWS)





# Data Lake - Aquisição de Dados em Batch

An abstract background diagram consisting of several colored circles (blue, green, yellow, orange) connected by thin lines, suggesting a network or data flow. A large green rounded rectangle is centered over this diagram.

Data Lake na Nuvem  
(Usando AWS)

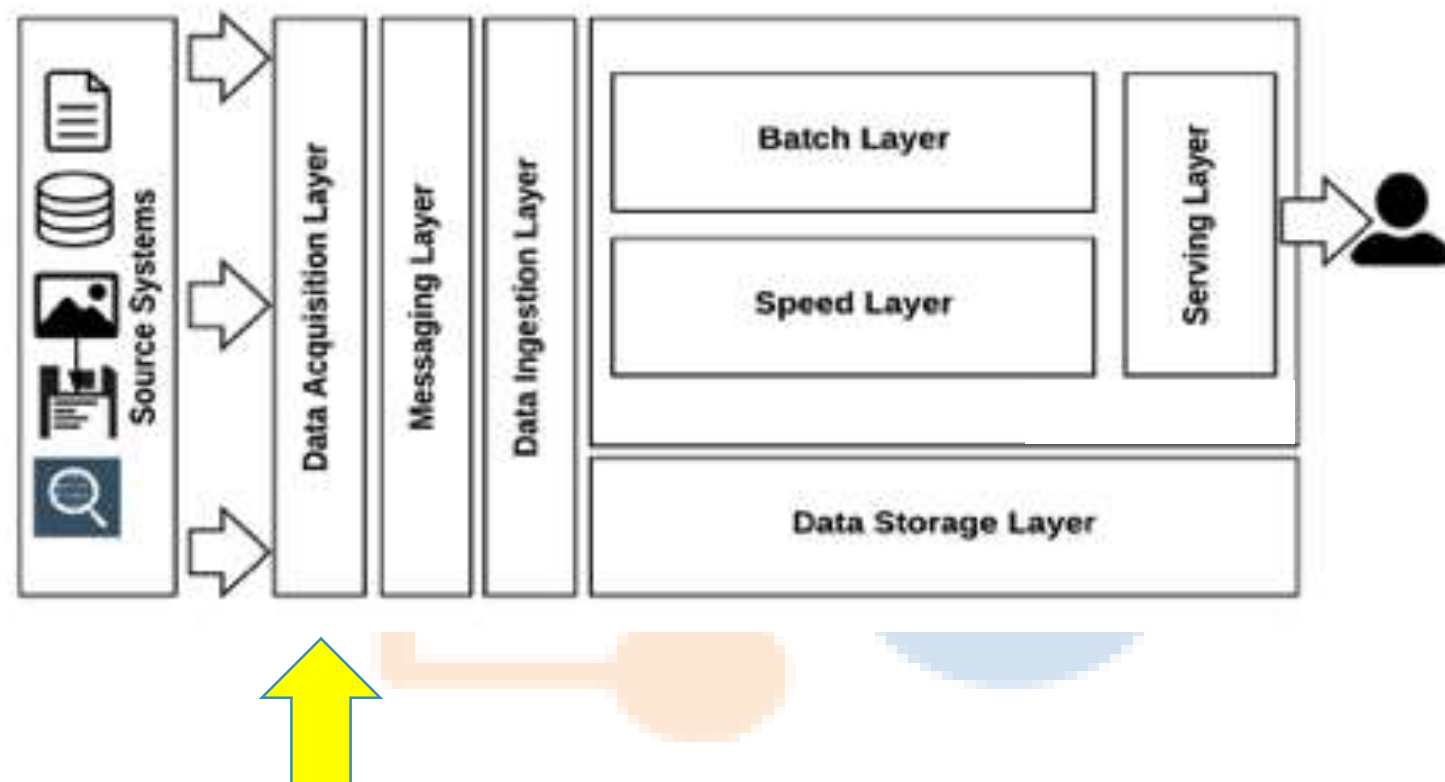




# Data Lake - Aquisição de Dados em Batch

Aquisição de  
Dados em  
Batch  
(Cap05)

Aquisição de  
Dados em  
Streaming  
(Cap06)





# Contexto no Data Lake

## Aquisição de Dados





# Contexto no Data Lake

## Aquisição de Dados

O processo de indução de dados de vários sistemas de origem é chamado de aquisição de dados. Em nosso Data Lake, temos uma camada definida (na verdade, a primeira) que tem exatamente essa responsabilidade.





# Contexto no Data Lake

## Aquisição de Dados



*Relational Database*



*Hadoop Distributed File System*







# O Que São Dados em Batch?





# O Que São Dados em Batch?

Processamento batch, no contexto de sistemas operacionais, é um termo referente a um processamento de dados que ocorre através de um lote de tarefas enfileiradas, de modo que o sistema operacional só processa a próxima tarefa após o término completo da tarefa anterior.





# O Que São Dados em Batch?

Processamento batch é utilizado quando do tratamento de arquivos que contêm diversos registros, também conhecido como processamento por lote de dados. Este recurso é muito utilizado em sistemas distribuídos onde as aplicações de front-end tem restrições, seja de custos, de telecomunicações ou de arquitetura, para efetuar atualizações online/real-time nos sistemas centralizados.





# O Que São Dados em Batch?

Opta-se por alimentar um banco de dados local durante o dia e envio de um arquivo sequencial contendo todos os registros para atualizações no período noturno nos sistemas centrais. **Esses são os dados em batch!**





# O Que São Dados em Batch?

O processamento batch tem a característica de não exigir a interação do usuário com a aplicação. Todas as entradas e saídas de dados da aplicação são implementadas por algum tipo de memória secundária, geralmente arquivos em disco. Alguns exemplos de aplicações originalmente processadas em batch são programas envolvendo cálculos numéricos, compilações, ordenações, backups e todos aqueles onde não é necessária a interação com o usuário.



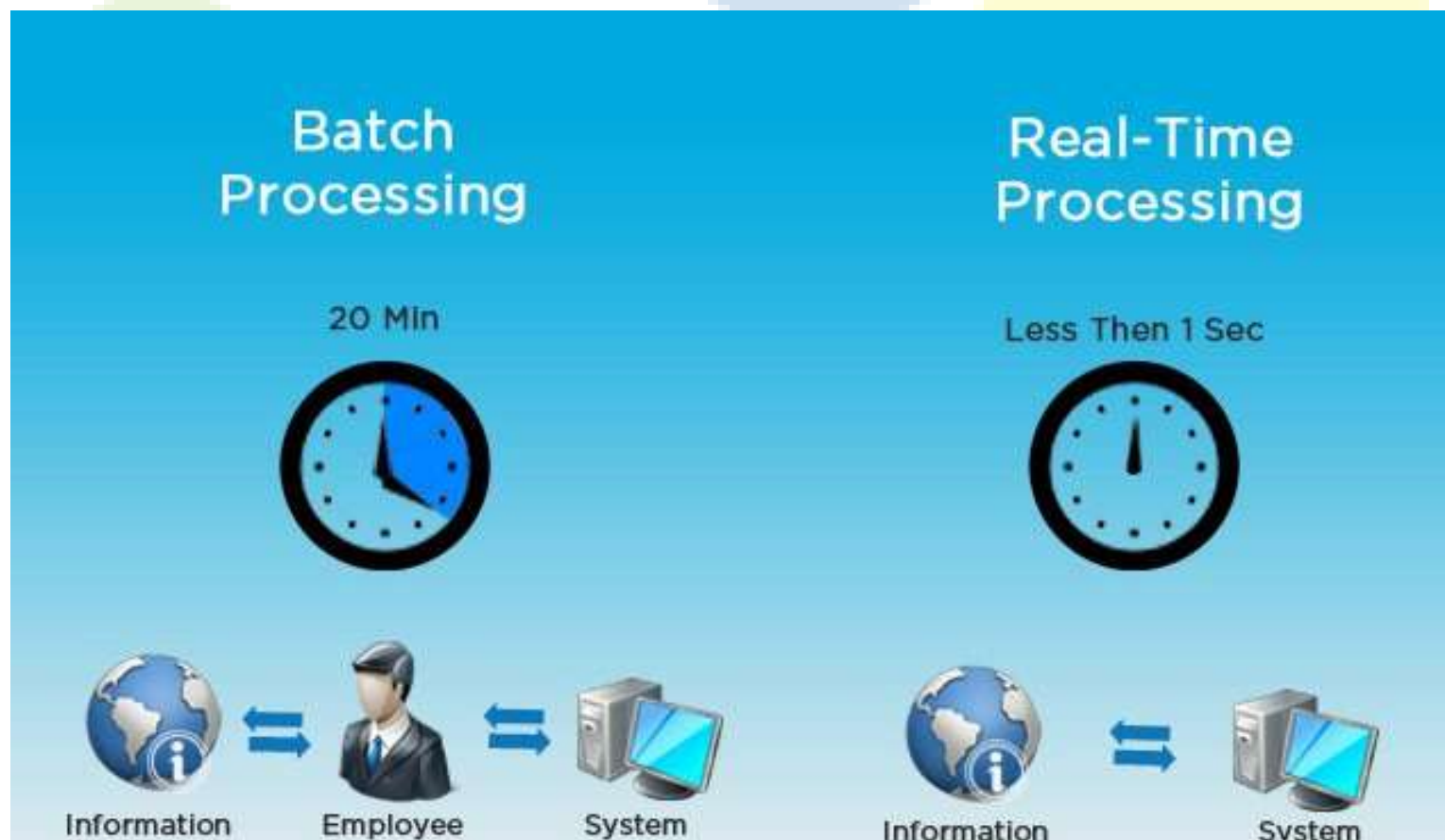


# Real-Time Data x Batch Data





# Real-Time Data x Batch Data





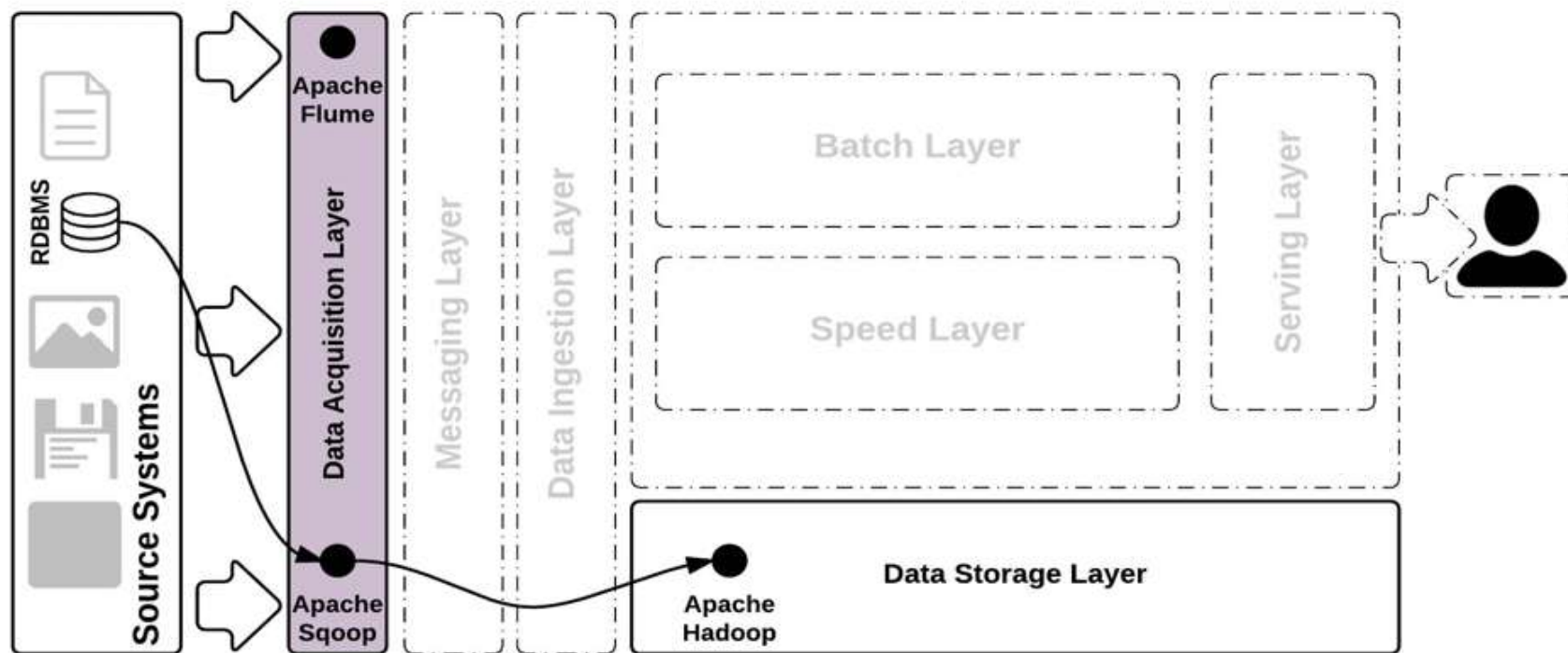


# Mapeamento de Dados





# Mapeamento de Dados

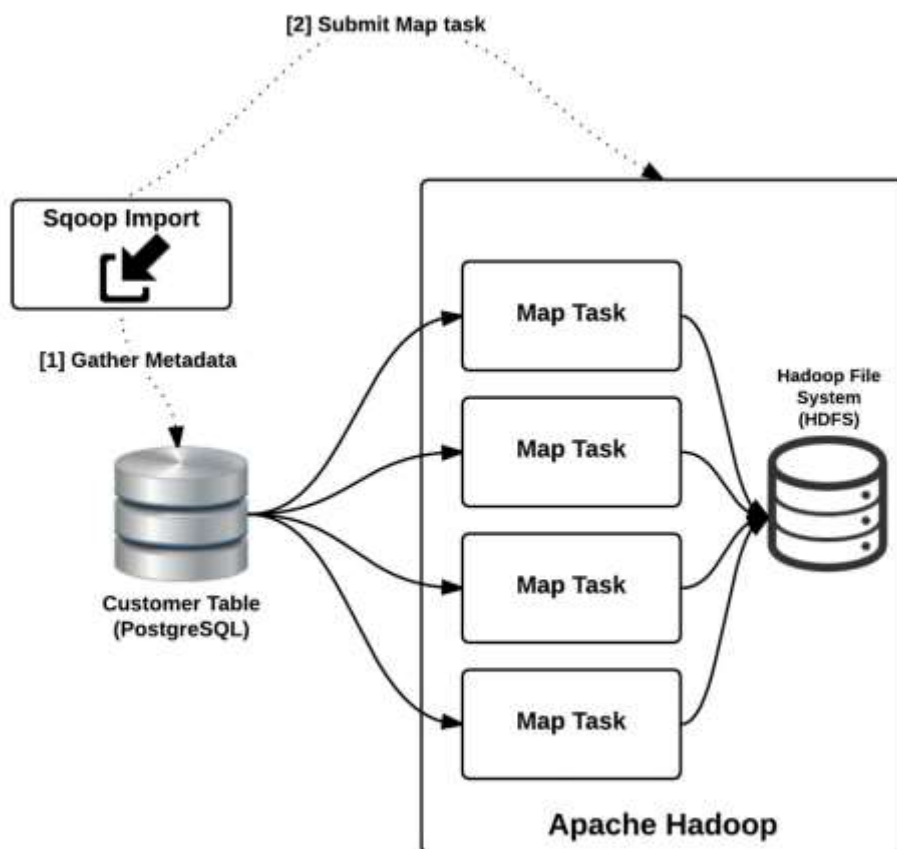




# Arquitetura de Importação/Exportação de Dados com Sqoop



# Arquitetura de Importação/Exportação de Dados com Sqoop

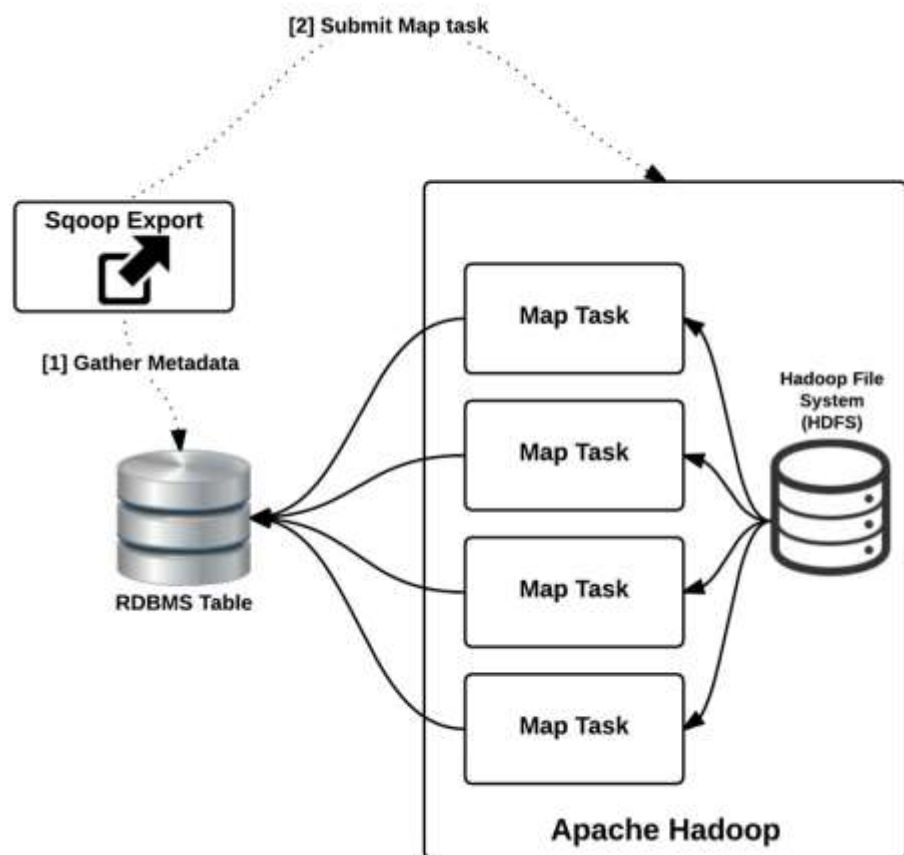


Importação





# Arquitetura de Importação/Exportação de Dados com Sqoop



Exportação





# Conectores Sqoop





# Conectores Sqoop





# Muito Obrigado.

É um prazer ter você aqui.  
Tenha uma excelente jornada de aprendizagem.

