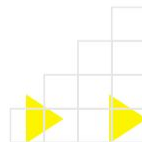




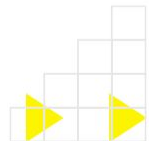
Curso 3: Administração de Dados Complexos em Larga Escala -- Apache Kylin --

Prof. Jose Fernando Rodrigues Junior

Objetivo: apresentar a solução OLAP Apache Kylin



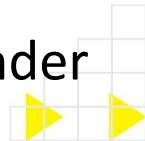
Apache Kylin: Extreme OLAP Engine for Big Data



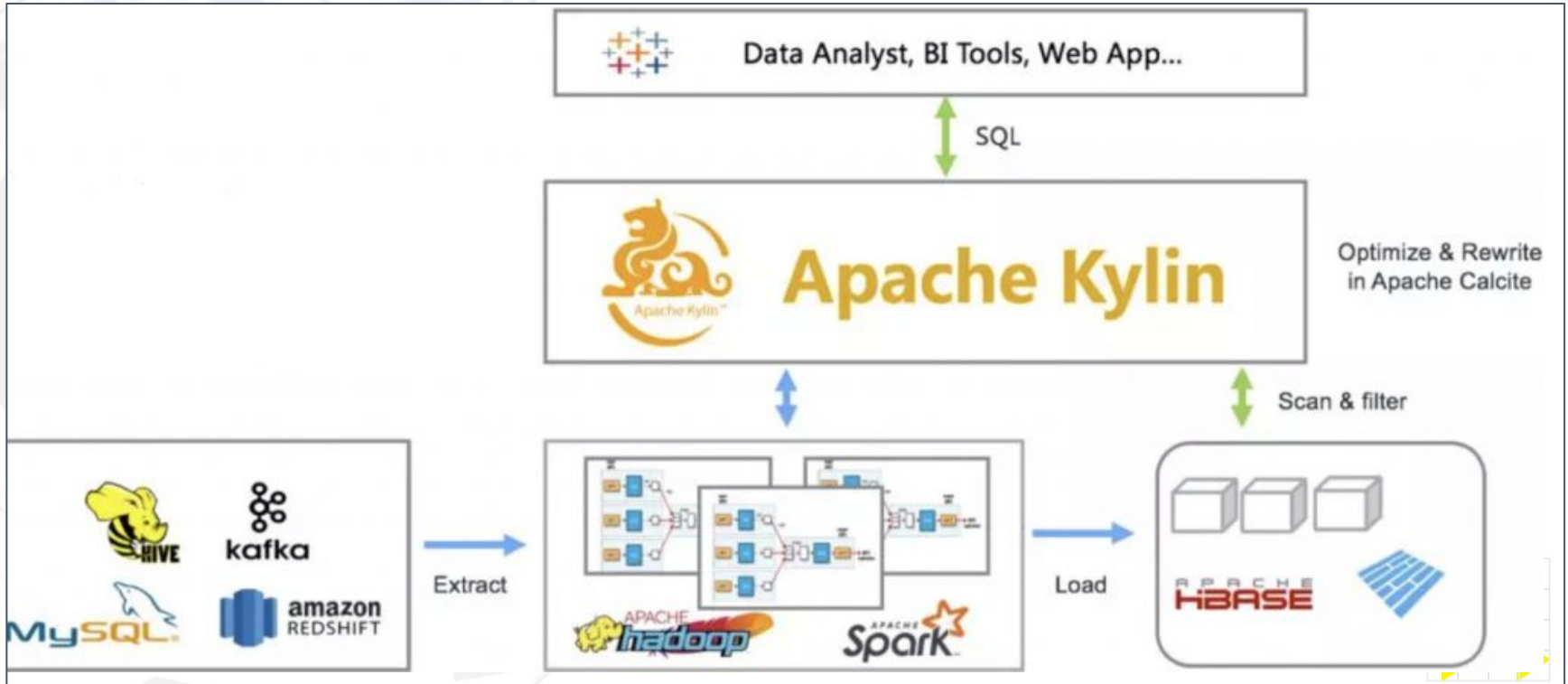


Apache Kylin: Extreme OLAP Engine for Big Data

- Projeto originário na empresa **eBay**, em **2013**; open-sourced em 2014;
- Projetado sobre o ecossistema **Hadoop**; também funciona sobre **Spark**;
⇒ Para saber mais: [Hadoop x Spark](#)
- Alto desempenho em escala **Big Data**;
- Faz uso de técnicas de **pré-computação** de queries para responder consultas OLAP em segundos.

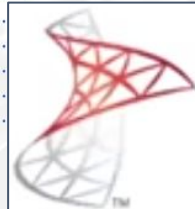


Apache Kylin





Outras soluções OLAP



Microsoft®
SQL Server®
Analysis Services

ORACLE®

ORACLE ESSBASE

IBM
COGNOS®

MicroStrategy®

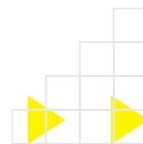
⇒ O Apache Kylin foi escolhido por ser open-source e por ter ampla aceitação no mercado.





Cenário

- ❏ Como visto, ao suportar SQL e grandes volumes de dados, **o Hive é teoricamente capaz de realizar OLAP;**
- ❏ No entanto, ele **não o faz de maneira tão eficiente quanto o Kylin;**
- ❏ Há **requisitos que não são cobertos pelo Hive.**





Requisitos para Big Data OLAP



Requisitos **de uso**:



queries sobre **bilhões de tuplas** executadas em segundos;



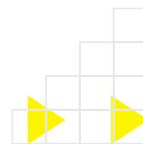
Ansi SQL para analistas e engenheiros;



abstração **OLAP**;



integração fácil com ferramentas de análise: Tableau, Spotfire, PowerBI, Saas, Excel, ...

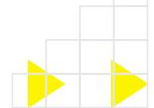


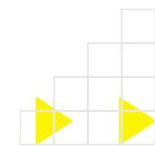
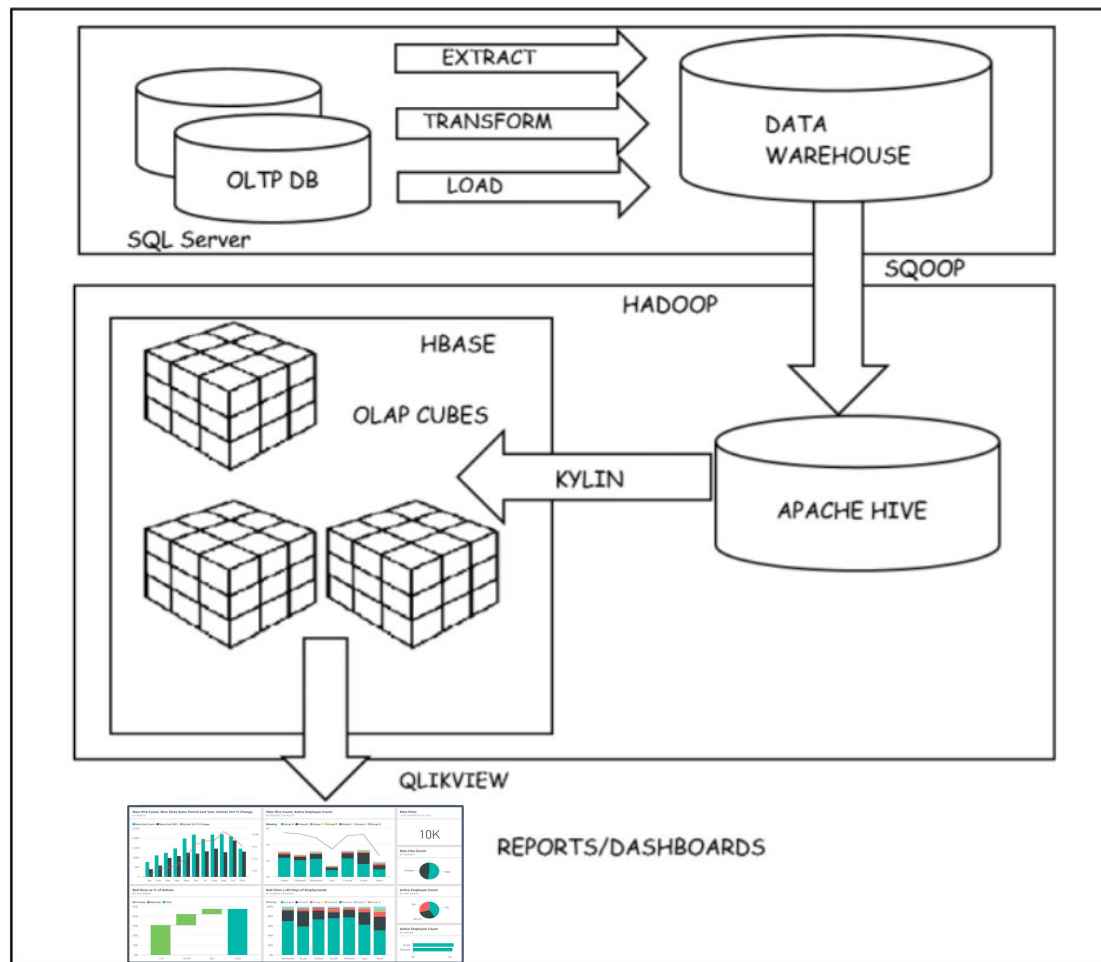
Apache Kylin



Solução OLAP baseada no ecossistema Apache Hadoop:

- **Sqoop**: importa dados relacionais para tabelas Hive;
- **Hive (DW)**: armazenagem, e disparo de jobs MapReduce;
- **MapReduce**: processamento distribuído escalável e abstrato;
- **HBase DB**: organiza e recupera os cubos pré-computados;
- **HDFS**: dados de maneira distribuída para Hive e HBase;
- **Calcite**: interpretador de consultas SQL.

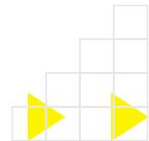




Apache HBase



- Uma base de dados **NoSQL** - <https://db-engines.com/en/ranking>;
- Funciona sobre o sistema de arquivos **HDFS**, de modo distribuído;
- Provê acesso aleatório aos dados, **com baixa latência**, via hashing;
- Funciona como um **hashmap persistente**;
- Derivado do **Bigtable** do Google;
- **Column-oriented**: menos espaço, análise mais rápida;
- Organizado na forma de famílias de colunas, versionadas por *timestamp*.



Apache HBase



Mas porque o Kylin precisa usar o HBase se ele já usa o Hive?

R.: o Hive tem uma grande latência pois depende de *jobs* MapReduce, o que faz com que ele seja rápido apenas na recuperação de grandes quantidades de dados em sequência; além disso, ele não suporta update de dados individuais, apenas de arquivos inteiros (devido ao HDFS).

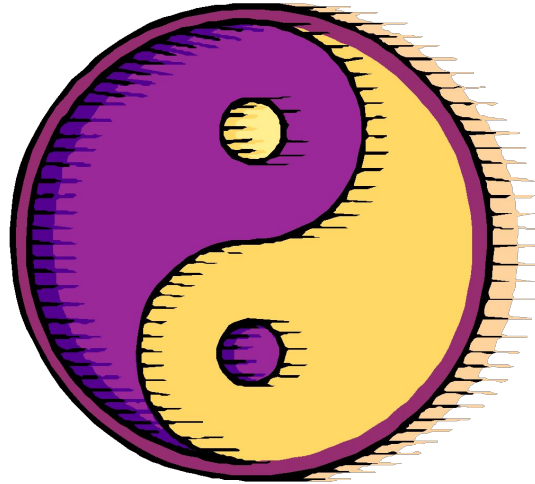
"O Hive não foi projetado para processamento de transações e não oferece consultas em tempo real e atualizações de nível de linha. É melhor usado para trabalhos em lote em grandes conjuntos de dados imutáveis (como logs da web)."

⇒ Assim, o HBase supre às necessidades de armazenamento/acesso da pré-computação de agregações do Kylin.

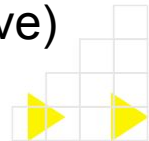
Juntos data warehouse e bancos de dados provém uma solução completa



Bancos de dados
Inserção/Atualização
(HBase)









Data Warehouse
Acesso aos dados
(OLAP - Hive)

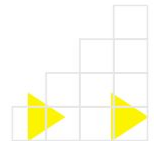


Apache Kylin



Características:

-  **Consultas interativas** em segundos;
-  **Multidimensional Cube:** os analistas podem definir o modelo de dados e pré-computar o cubo;
-  **OLAP engine** rápido e escalável: baseado em pré-computação de cubos;
-  **Compressão** e **refresh** dos cubos;
-  **Interface Web;**
-  **Integração com BI:** integração com QlikView, Tableau, PowerBI ou Excel.



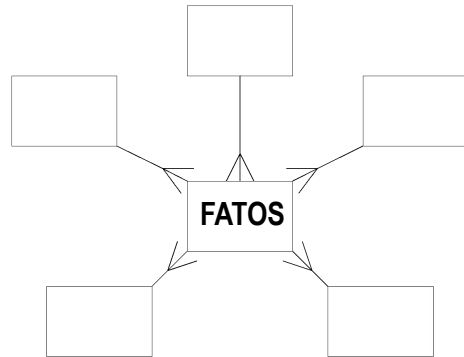
Apache Kylin Global Adoptions



Apache Kylin

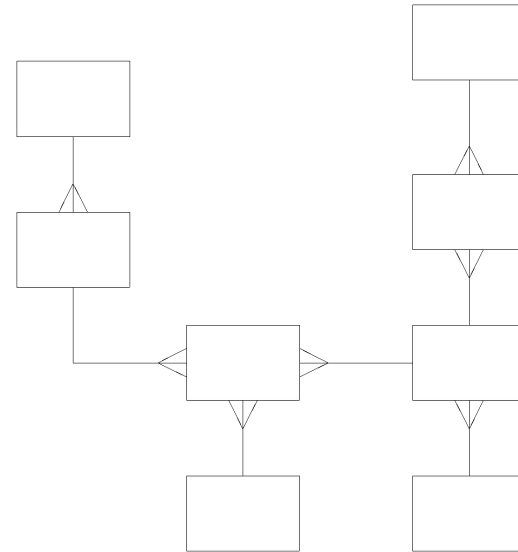


Apache Kylin

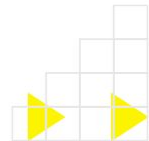


Esquema
estrela

Banco de dados operacional



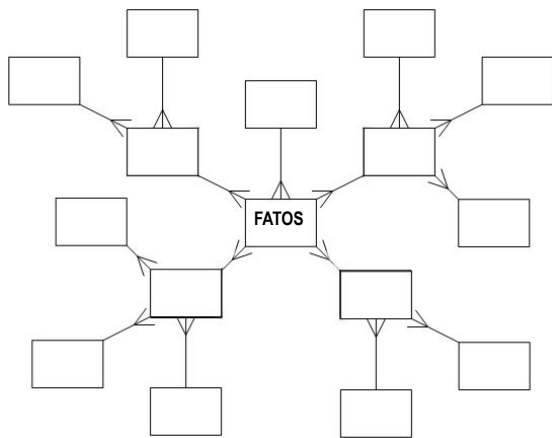
Esquema complexo





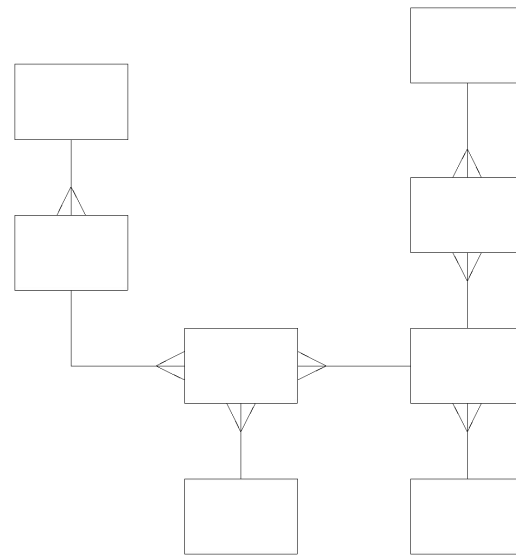
Apache Kylin

Apache Kylin

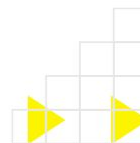


Esquema floco de neve
(snow flake)

Banco de dados operacional



Esquema complexo

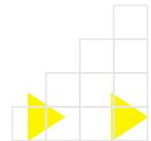


Apache Kylin



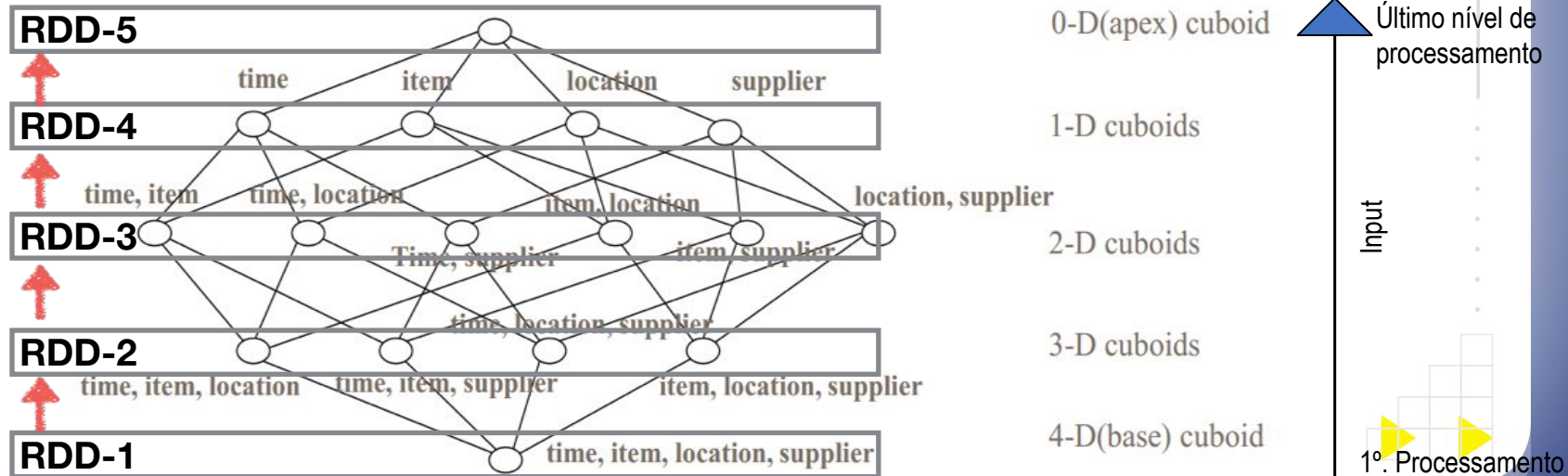
Note que o esquema da base de dados operacional **raramente** (quase nunca) tem estrutura estrela ou floco de neve.

Mas estas estruturas (estrela ou floco de neve) precisam ser definidas para fins de **(pré) processamento** via interface de design do sistema OLAP. Geralmente, uma sub estrutura da base operacional.



Apache Kylin - Desempenho

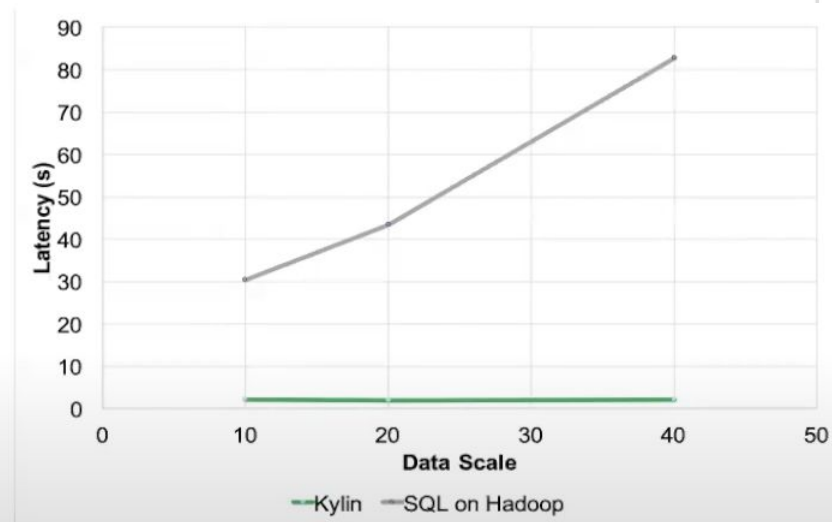
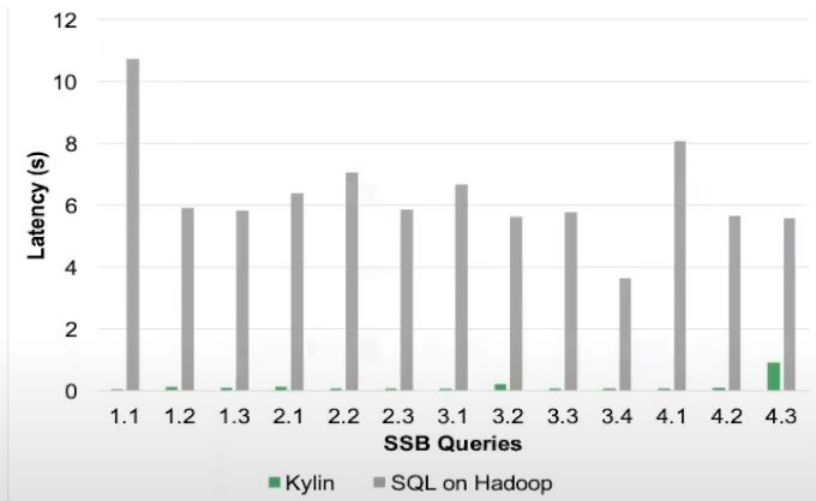
- O Apache Kylin usa o algoritmo **Layered Cubing**;
- **Pré-computação da agregação** considerando a **granularidade mais fina de todas as dimensões**;
- A seguir, qualquer sub agregação pode ser computada tendo como **input o resultado da agregação com mais dimensões**.





Apache Kylin - Desempenho

- Com o algoritmo de *Layered Cubing*, todas as agregações são **pré-computadas**;
- A interação com o cubo passa a ser baseada em resultados pré-computados.



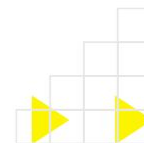


Apache Kylin - Desempenho

Query Performance -- Compare to Hive



#	Query Type	Return Dataset	Query On Kylin (s)	Query On Hive (s)	Comments
1	High Level Aggregation	4	0.129	157.437	1,217 times
2	Analysis Query	22,669	1.615	109.206	68 times
3	Drill Down to Detail	325,029	12.058	113.123	9 times
4	Drill Down to Detail	524,780	22.42	6383.21	278 times
5	Data Dump	972,002	49.054	N/A	



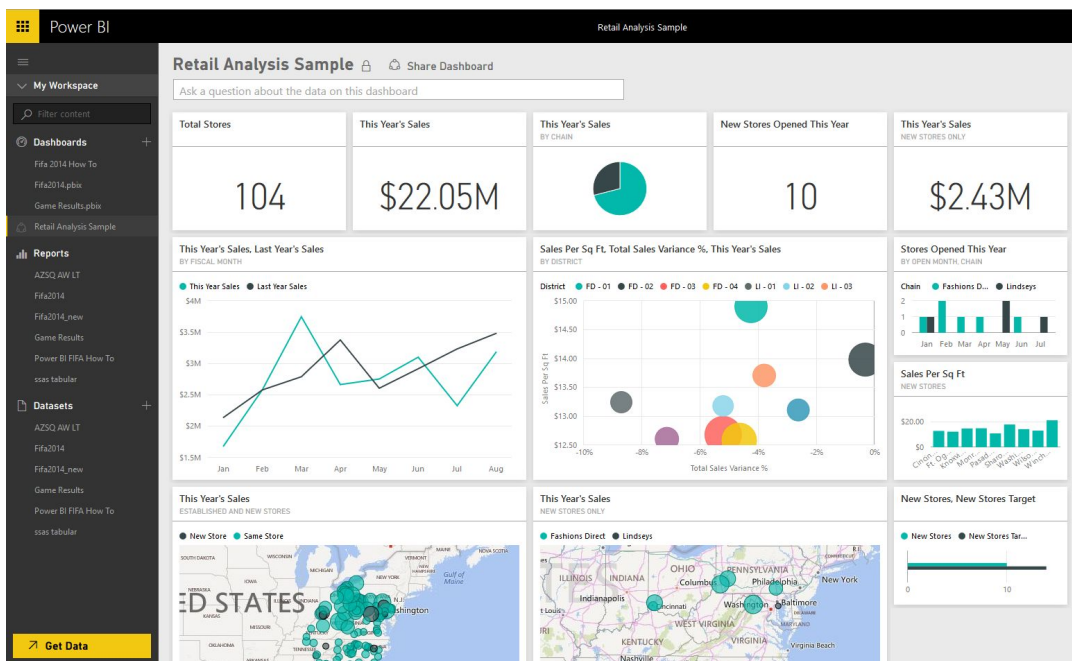


Apache Kylin - Integração



Integração com BI: integração com PowerBI ou Excel.

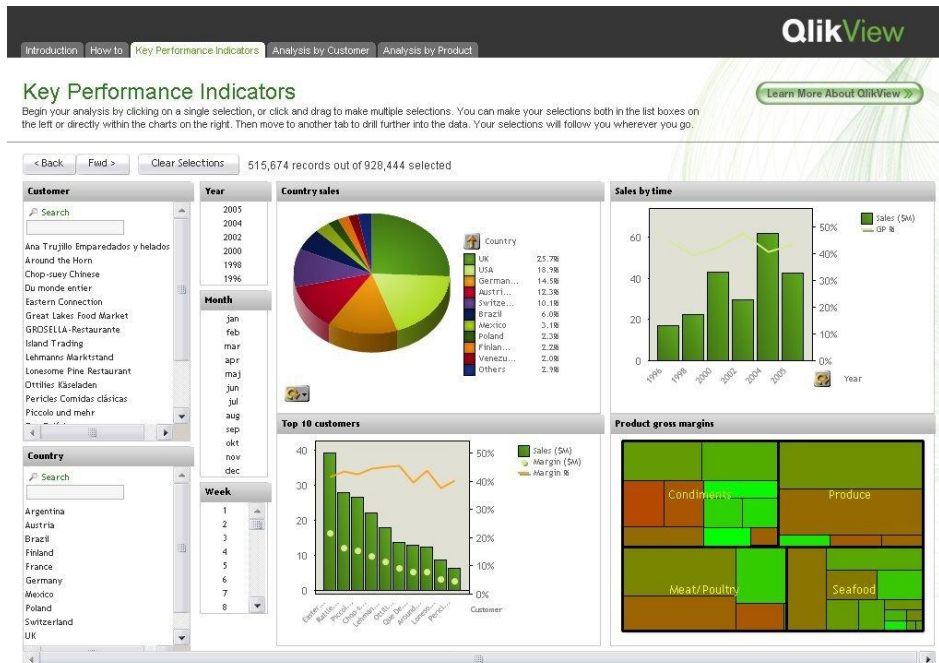
QlikView, Tableau,





Apache Kylin - Integração

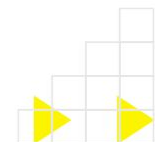
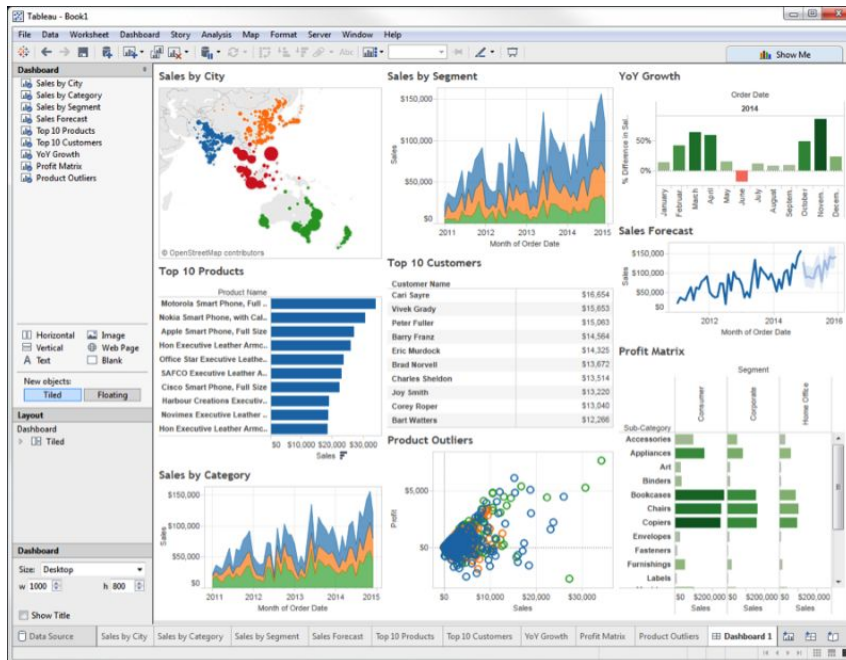
Integração com BI: integração com QlikView, Tableau, PowerBI ou Excel.





Apache Kylin - Integração

 **Integração com BI:** integração com QlikView, Tableau, PowerBI ou Excel.

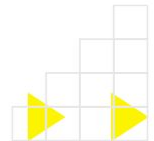


Apache Kylin



O Kylin oferece um Wizard Web para a definição de uma análise de dados OLAP por meio de 4 passos:

- 1) Criação do projeto;
- 2) Carregar dados do Hive;
- 3) Criação do modelo estrela (ou floco de neve);
- 4) Criação do cubo.

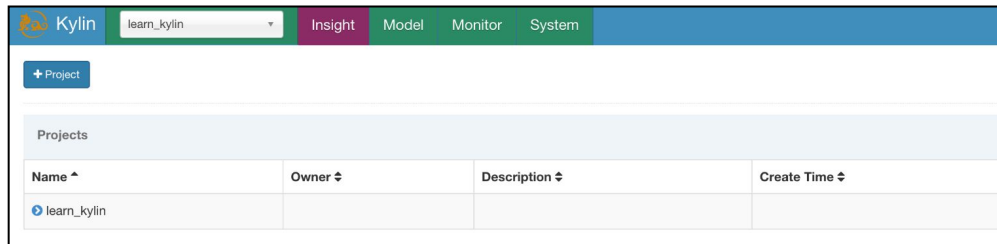




Apache Kylin

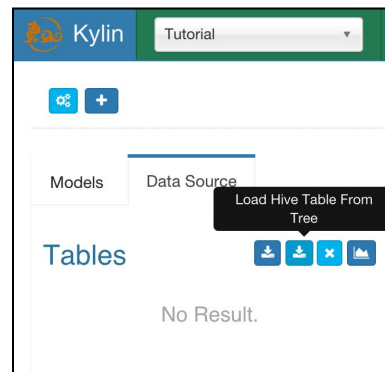
O Kylin oferece um Wizard Web para a definição de uma análise de dados OLAP por meio de 4 passos:

1) Criação do projeto:



Name ^	Owner ⇅	Description ⇅	Create Time ⇅
learn_kylin			

2) Carregamento de dados do Hive:





Apache Kylin

3) Criação do modelo de dados, estrela ou floco de neve:

Model Designer

Model Info Data Model Dimensions Measures Settings

Select your measures

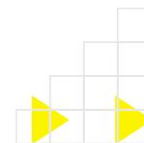
PRICE X ITEM_COUNT X SELLER_ID X

4) Criação do cubo:

Cube Designer

Cube Info Dimensions Measures Refresh Setting Advanced Setting Overview

Model Name	Tutorial_Model	Description
Cube Name	Tutorial_Cube	
Fact Table	DEFAULT.KYLIN_SALES	
Lookup Table	1	
Dimensions	5	
Measures	6	



Apache Kylin



Exercício *Hands on* – definindo um cubo com o assistente do Apache Kylin:

- 1) Use um docker-container com o Kylin pré-configurado:
http://kylin.apache.org/docs30/install/kylin_docker.html*
*Recomenda-se o uso de máquinas com 16 GB de memória
- 2) Seguir o roteiro:
http://kylin.apache.org/docs20/tutorial/create_cube.html
 - Default username/passwd: ADMIN/KYLIN
 - Esquema do exemplo: http://kylin.apache.org/docs/howto/sample_dataset.html

Demonstração completa: vídeo Aula08