Big Data e Data Warehousing

Amélia Pessoa

Agenda

Introdução

Data warehouses convencionais

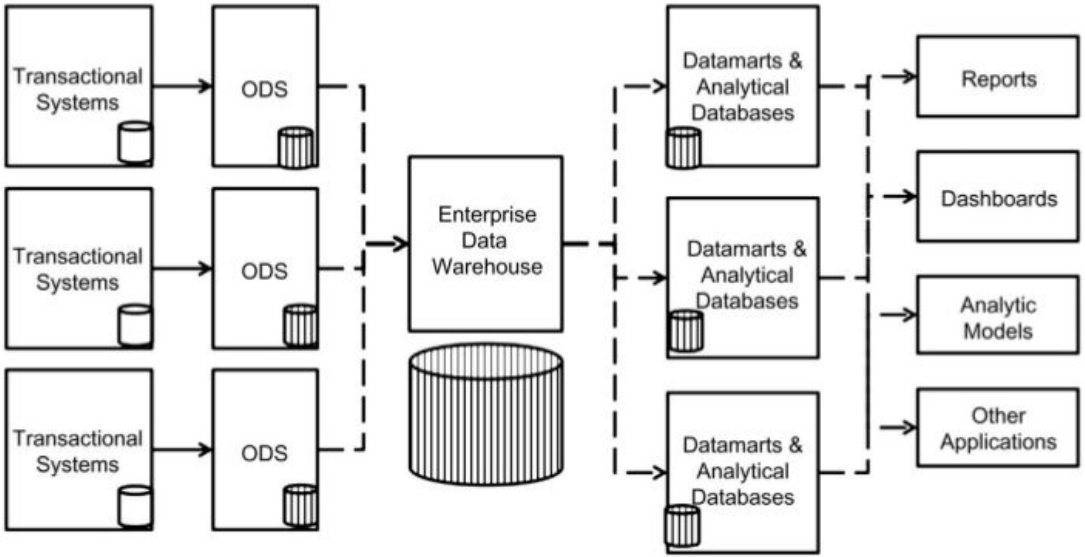
Novos data warehouses

- Desafios e componentes
- Integração de dados
- Modelos de arquitetura
- Soluções
- Estrutura Semântica

Conclusões

Introdução

Data warehouses convencionais



Fragmentação de dados

CPU subutilizada

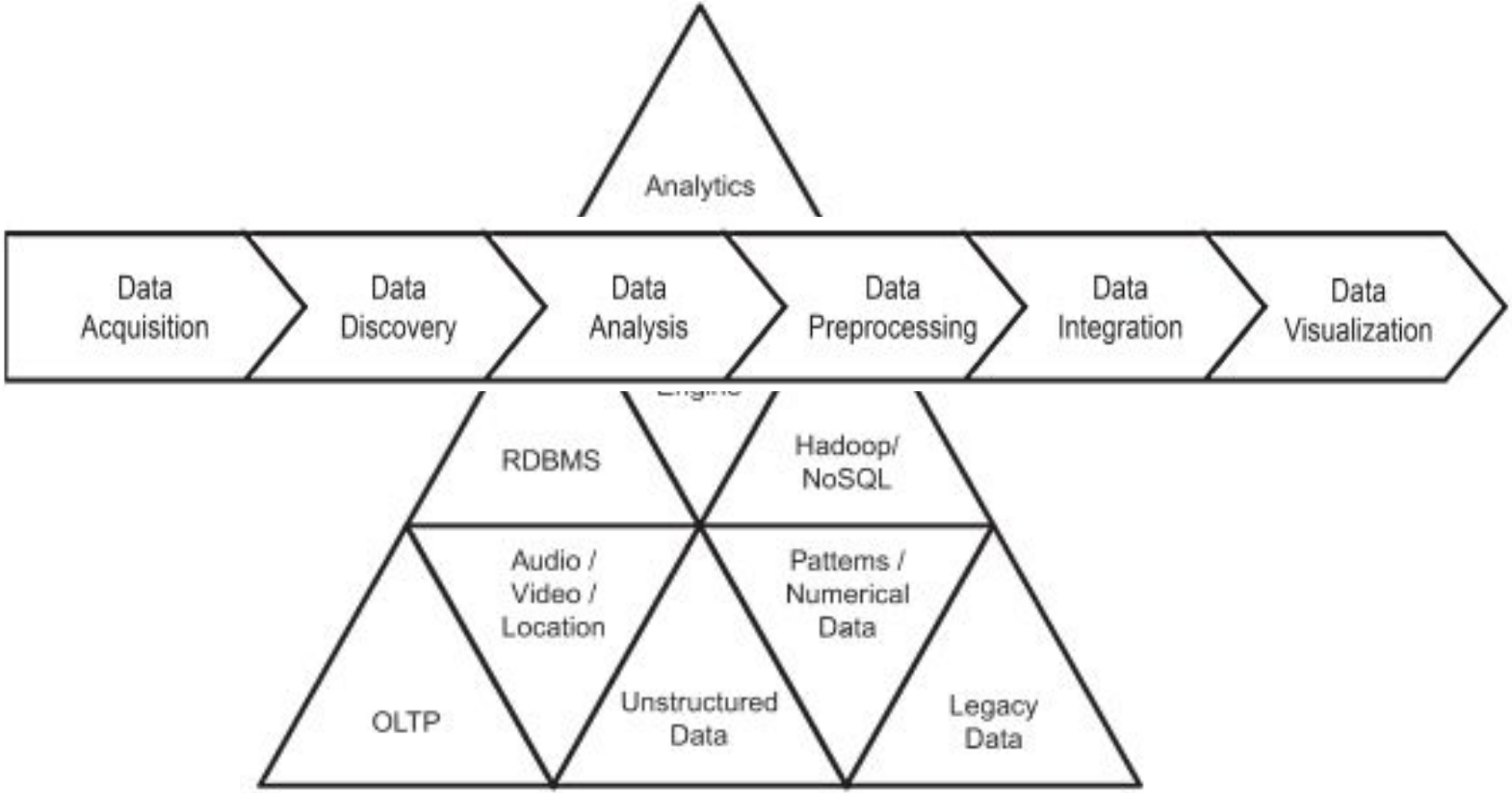
Memória subutilizada

Consultas deficientes

Dados do novo data warehouse

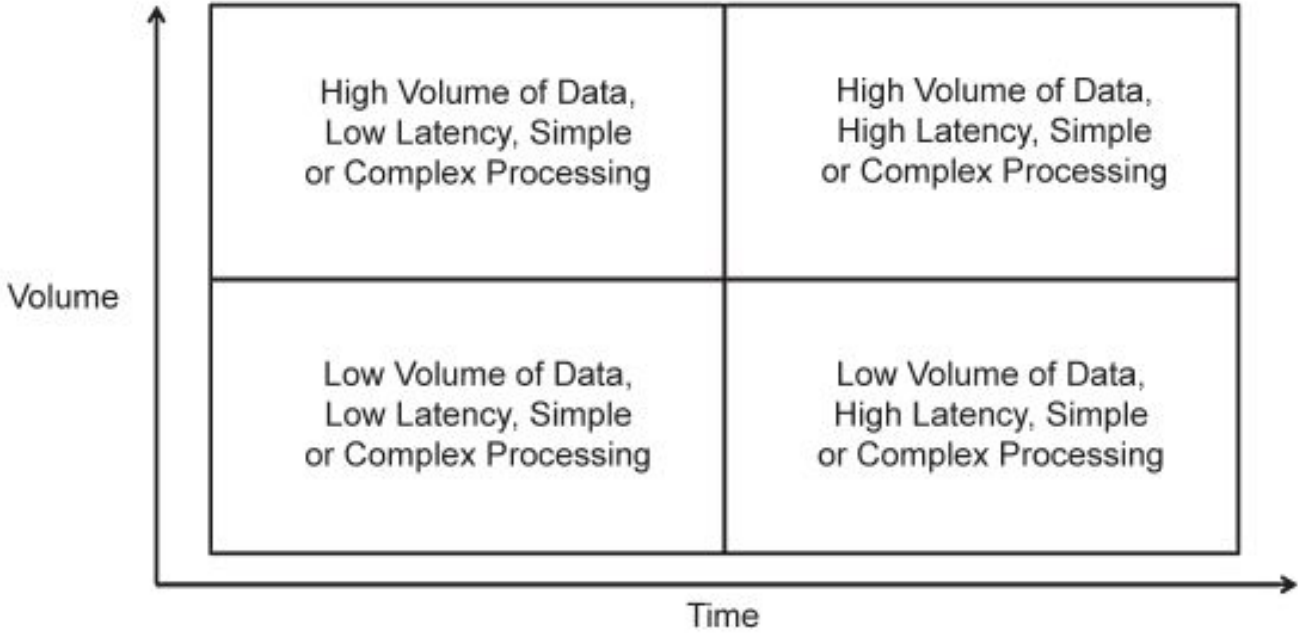
- Não têm uma arquitetura finita e podem ter vários formatos.
- Não são auto-suficientes e precisam de várias regras de negócios externas para processá-los.
- Precisam de mais processamento analítico.
- Dependem de metadados para criar contexto.
- Não têm especificidade com volume ou complexidade.
- São semi-estruturados ou não estruturados.
- Precisam de múltiplos ciclos de processamento.
- Precisam de mais governança do que os dados no banco de dados.
- Não têm qualidade definida.

Componentes do novo data warehouses



Integração de dados

Carga de trabalho



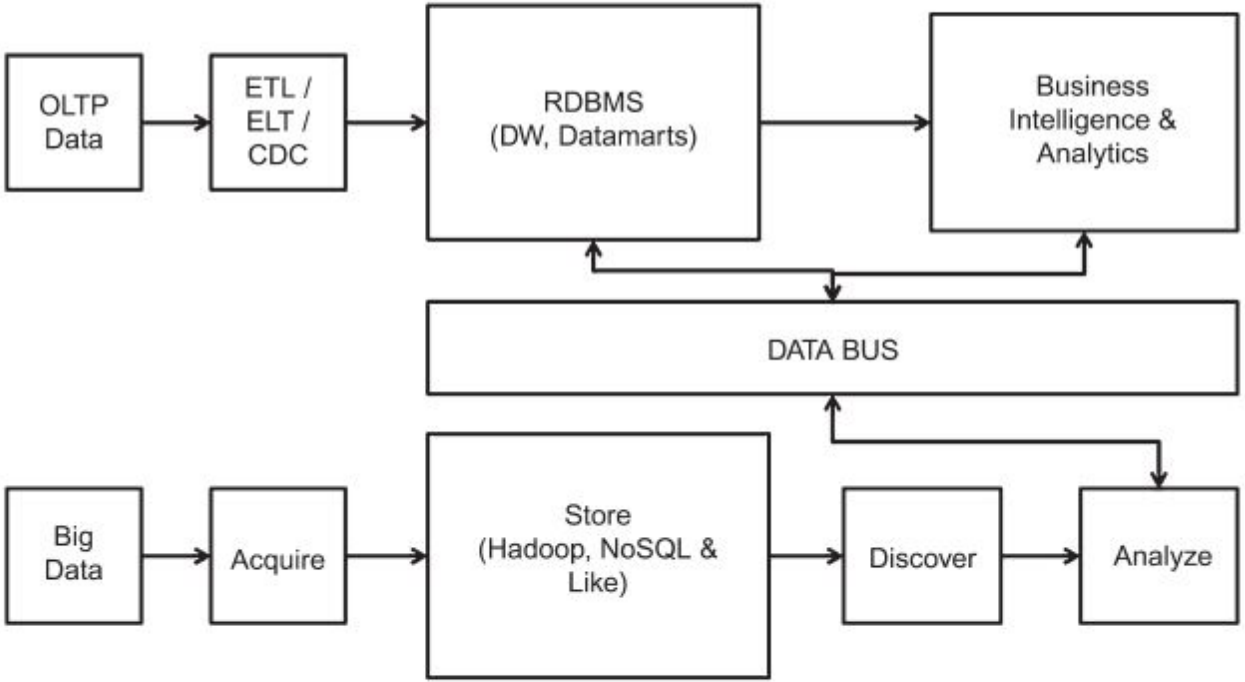
Integração de dados

Arquitetura Física

- Carga
- Disponibilidade
- Volume
- Desempenho de armazenamento
- Custos Operacionais

Modelos de Arquiteturas

Integração externa de dados



Modelos de Arquiteturas

Integração externa de dados

Prós:

- Projeto escalável para bancos de dados relacionais e Big Data.
- Redução da sobrecarga no processamento.
- A complexidade do processamento pode ser isolada através da aquisição de dados, limpeza de dados, descoberta e integração de dados.
- Arquitetura modular de integração de dados.
- Implementação de arquitetura física heterogênea, oferecendo a melhor integração com a camada de processamento de dados.

Modelos de Arquiteturas

Integração externa de dados

Contras:

- Arquitetura de barramento de dados pode se tornar cada vez mais complexa.
- Arquitetura de metadados pode se tornar deficiente devido a várias camadas de processamento de dados.
- A integração de dados pode se tornar um gargalo de desempenho durante um período de tempo.

Modelos de Arquiteturas

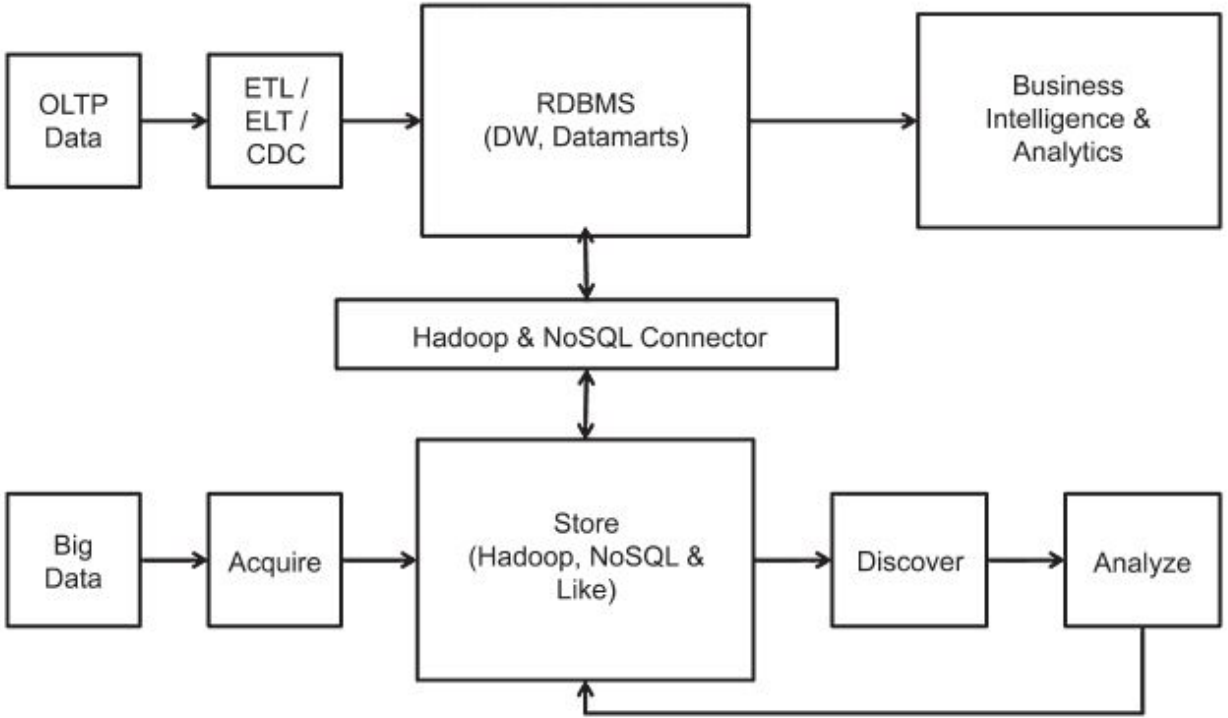
Integração externa de dados

Evitar:

- Demasiada complexidade de dados em qualquer camada de processamento.
- Metadados pobres.
- Análise incorreta de dados nas camadas Big Data.
- Níveis de integração incorretos (na granularidade de dados) dentro das camadas Big Data.
- Aplicação incorreta da integração do barramento de dados.

Modelos de Arquiteturas

Hadoop e RDBMS



Modelos de Arquiteturas

Hadoop e RDBMS

Prós:

- Projeto escalável para bancos relacionais e Big Data.
- Arquitetura modular de integração de dados.
- Implementação de arquitetura física heterogênea, oferecendo a melhor integração com a camada de processamento de dados.
- As soluções de metadados podem ser alavancadas com relativa facilidade em toda a solução.

Modelos de Arquiteturas

Hadoop e RDBMS

Contras:

- O desempenho do conector Big Data é a maior área de fraqueza.
- A integração de dados e a escalabilidade de consultas podem se tornar complexas.

Modelos de Arquiteturas

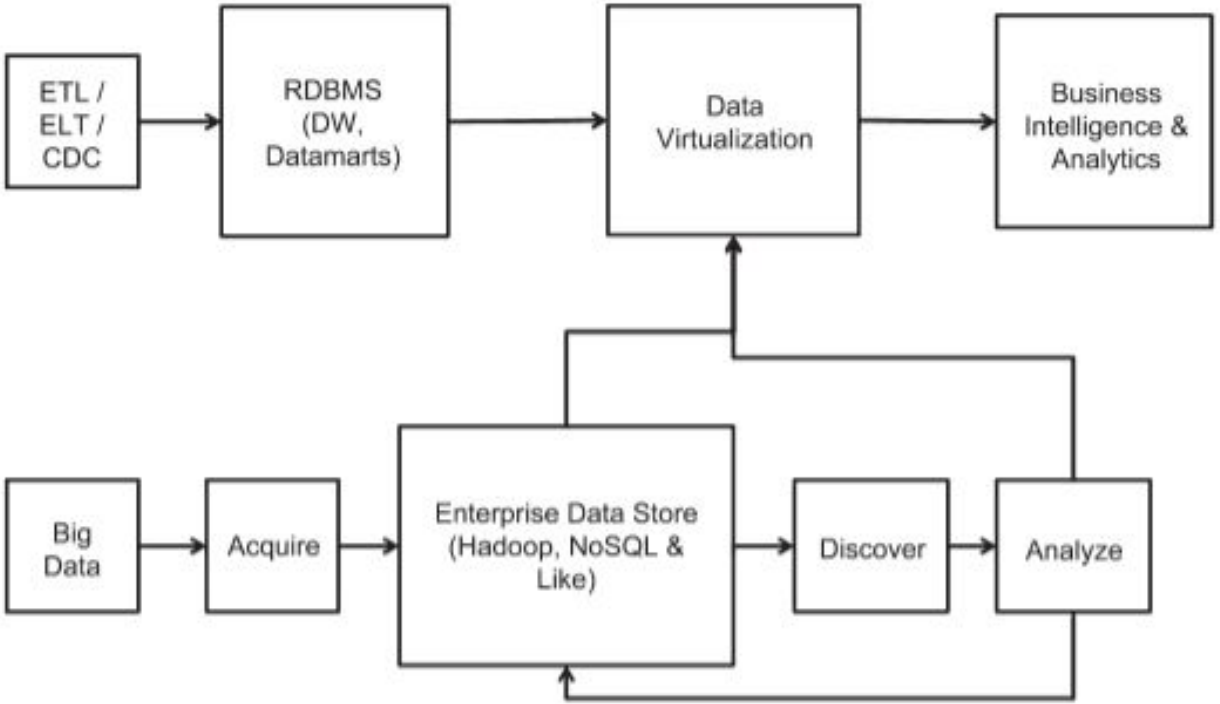
Hadoop e RDBMS

Evitar:

- Demasiada complexidade de dados em qualquer camada de processamento.
- Execução de grandes trocas de dados entre as diferentes camadas.
- Níveis de integração incorretos (na granularidade de dados).
- Aplicação de demasiadas complexidades de transformação usando os conectores.

Modelos de Arquiteturas

Virtualização de Dados



Modelos de Arquiteturas

Virtualização de Dados

Prós:

- Arquitetura extremamente escalável e flexível.
- Carga de trabalho otimizada.
- Fácil de manter.
- Menor custo inicial de implantação.

Modelos de Arquiteturas

Virtualização de Dados

Contras:

- A falta de governança pode criar muitos silos e degradar o desempenho.
- O processamento complexo de consultas pode se tornar degradado ao longo de um período de tempo.
- O desempenho na camada de integração pode precisar de manutenção periódica.

Modelos de Arquiteturas

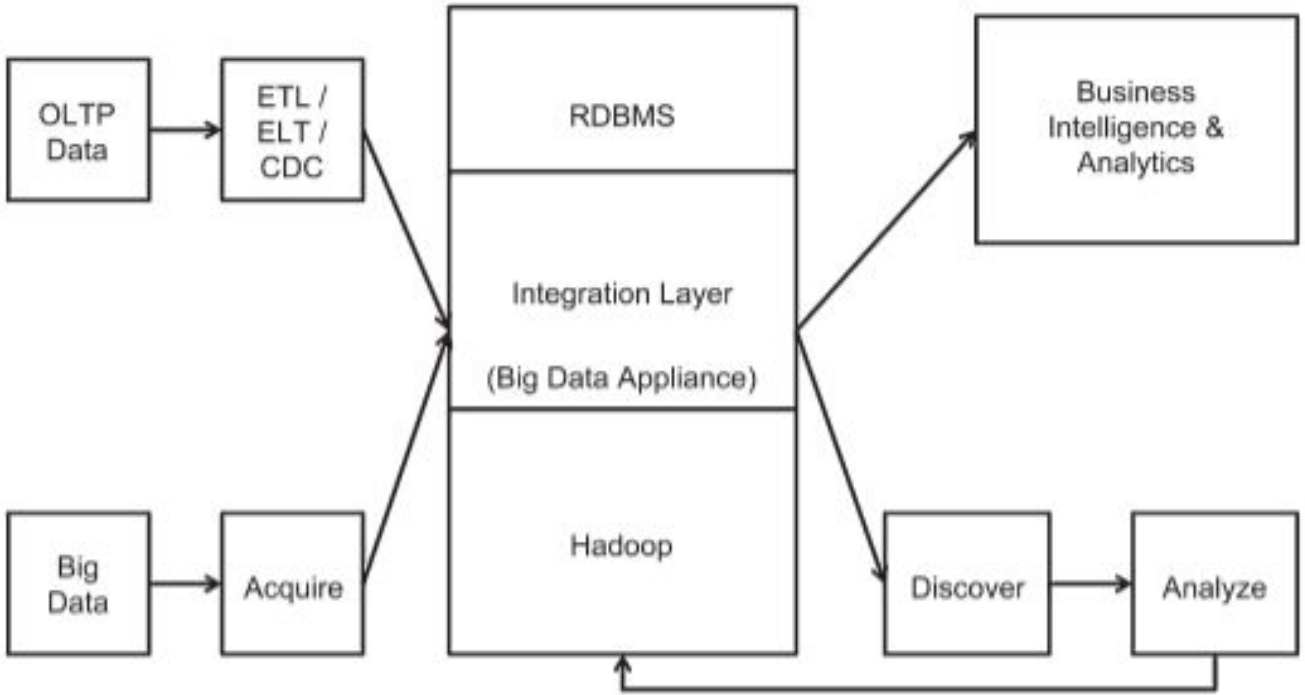
Virtualização de Dados

Evitar:

- Integração de dados acoplada de forma flexível.
- Granularidade de dados incorreta nos diferentes sistemas.
- Metadados deficientes em todos os sistemas.
- Falta de governança de dados.
- Integração de dados complexos envolvendo muitos cálculos na camada de integração.
- Arquitetura semântica mal projetada.

Modelos de Arquiteturas

Big Data Appliance



Modelos de Arquiteturas

Big Data Appliance

Prós:

- Design escalável e arquitetura modular de integração de dados.
- Implementação de arquitetura física heterogênea, oferecendo a melhor integração com a camada de processamento de dados.
- Personalizado e configurado para se adequar aos rigores de processamento, conforme exigido para cada organização.

Modelos de Arquiteturas

Big Data Appliance

Contras:

- Configuração personalizada é a maior fraqueza.
- A integração de dados e a escalabilidade de consultas podem se tornar complexas com o tempo.

Modelos de Arquiteturas

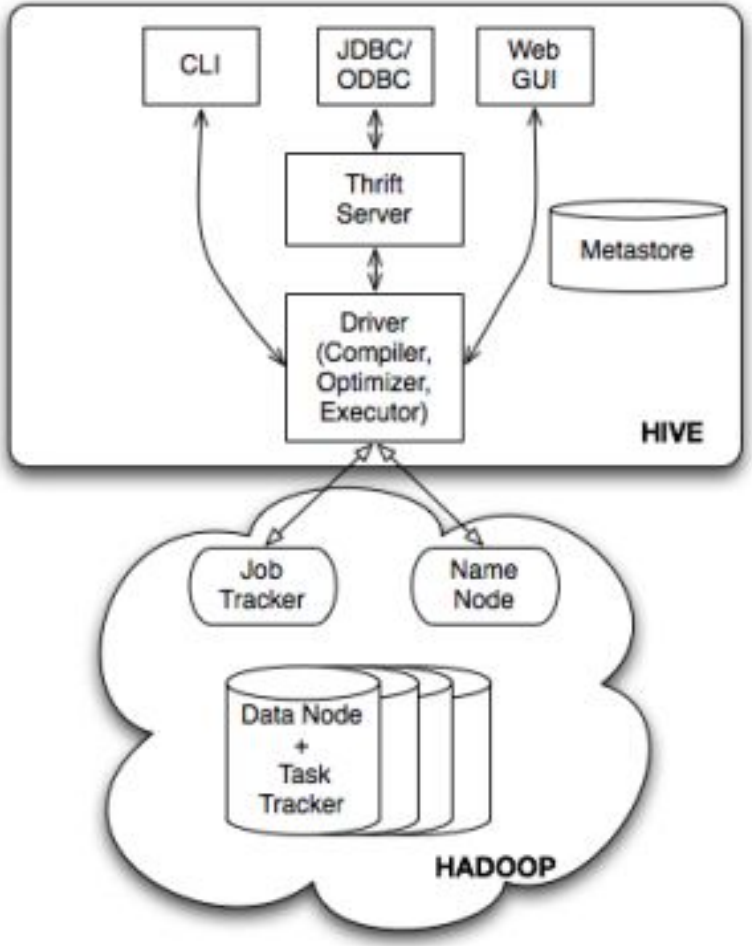
Big Data Appliance

Evitar:

- Configuração personalizada pode ser de manutenção pesada.
- Executar grandes trocas de dados entre as diferentes camadas pode causar problemas de desempenho.
- Dependência demais em qualquer camada de transformação cria gargalos de escalabilidade.
- Implementação de segurança de dados com integração LDAP deve ser evitada para as camadas não estruturadas.

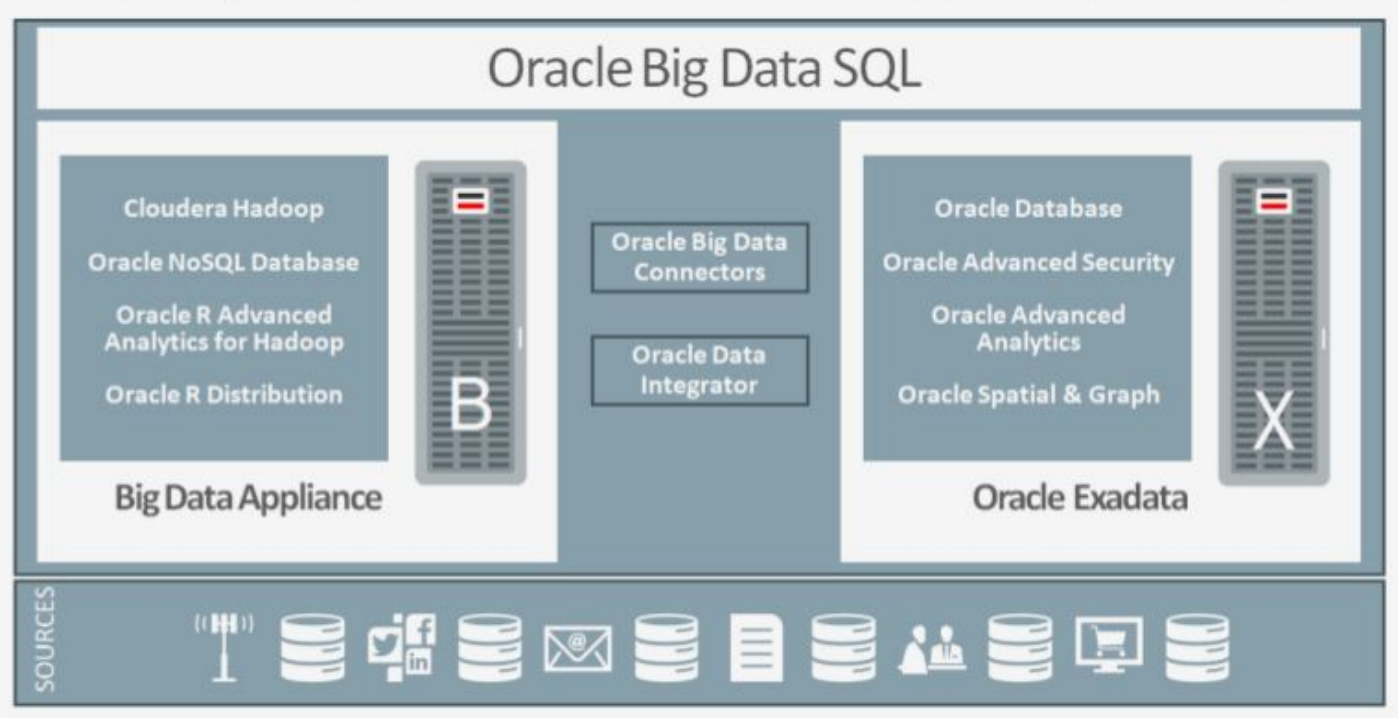
Soluções

Apache Hive



Soluções

Oracle Big Data Appliance

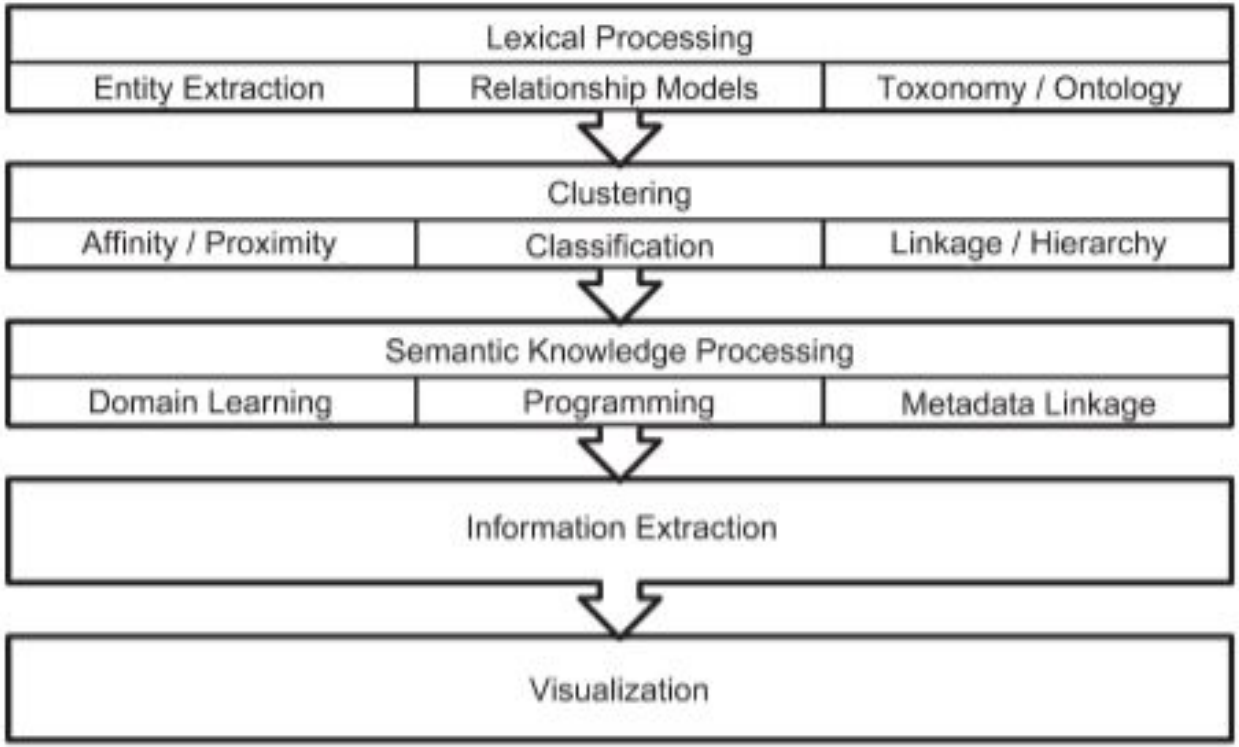


Soluções

Facebook

- Scribe
- HDFS / MapReduce
- Hive
- HiPal
- NoCron

Estrutura Semântica



Conclusões

- Desafios do Big Data (V’s);
- Integração de Soluções;
- Integração de Dados;
- Arquiteturas de data warehouses;
- Estrutura semântica.

Trabalhos futuros

a) Pesquisar soluções de Big Data mais utilizadas na indústria, comparando-as não só em termos de arquitetura assim como com estudos de caso;
b) Analisar as integrações de soluções no intuito de mitigar deficiências;
c) Aprofundar estudo no Apache Hive;
d) Aprofundar estudo na plataforma Hadoop (data warehouse);
e) Aprofundar estudo sobre data warehouses com repositórios NoSQL envolvendo suas modelagens e comparação de performance;
f) Avaliar ferramentas de Business Intelligence (BI) consolidadas no mercado no intuito de medir seus suportes ao Big Data;
g) Analisar o impacto da inserção de Big Data em ambientes de BI já em produção.

Referências

BAKSHI, Kapil. Considerations for big data: Architecture and approach. IEEE Aerospace Conference Proceedings, p. 1–7, 2012.

BRYANT, R; KATZ, Rh; LAZOWSKA, Ed. Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. Computing Research Association, p. 1–15, 2008. Disponível em: <http://www.just.edu.jo/~amerb/teaching/2-12-13/cs728/20123173012.pdf>.

DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: Simplied Data Processing on Large Clusters. Proceedings of 6th Symposium on Operating Systems Design and Implementation, p. 137–149, 2004.

DIJCKS, Jp. Oracle: Big data for the enterprise. Oracle White Paper, n. June, p. 16, 2012. Disponível em: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Oracle+:+Big+Data+for+the+Enterprise#0>.

DITTRICH, Jens; QUIAN, Jorge-arnulfo. Efficient Big Data Processing in Hadoop MapReduce. Proceedings of the VLDB Endowment, v. 5, n. 12, p. 2014–2015, 2012.

DOMO, Inc. Data Never Sleeps 3.0. Disponível em: <https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/>. Acesso em: 1 jun. 2016.

Referências

HUAI, Yin et al. Major Technical Advancements in Apache Hive. 2014.

KRISHNAN, Krish. Data Warehousing in the Age of Big Data. I ed. Waltham, MA, USA: Elsevier Inc, 2013.

MENON, Aravind. Big Data @ Facebook. MBDS '12: Proceedings of the 2012 workshop on Management of big data systems, p. 31, 2012.

MINELLI, Michael; CHAMBERS, Michelle; DHIRAJ, Ambiga. Big Data Analytics - Emerging BI and Analitics trends for today's businesses. [S.l: s.n.], 2013.

MOHANTY, Soumendra; JAGADEESH, Madhu; SRIVATSA, Harsha. Big Data Imperatives. I ed. New York, New York, USA: apress, 2013.

RUSSOM, Philip. Big data analytics . TWDI Best Practices Report, n. Fourth Quarter, p. 1–34, 2011.

THUSOO, Ashish et al. Hive - A Warehousing Solution Over a Map-Reduce Framework. Sort, v. 2, p. 1626–1629, 2009. Disponível em: <http://portal.acm.org/citation.cfm?id=1687609>.

Relatório

https://drive.google.com/open?id=0B9tZHMKCFZlHTkhwRmdMOHVjbDA