THE DEVELOPER'S CONFERENCE

Trilha Big Data

Data Lakes: da Teoria à Prática

Jorge Sandoval

M.Sc. em Computação Aplicada

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

- Termo cunhado em 2010 por James Dixon para distinguir entre a abordagem para gerenciar dados no Hadoop e *Data Marts* ou *Warehouses*
- "Se você pensa em um *Data Mart* como uma loja de água engarrafada - limpa, embalada e estruturada para fácil consumo - o *Data Lake* é um grande corpo de água em um estado mais natural." (Dixon, 2010).

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

- Local de armazenamento central em uma organização, empresa ou instituição
- Qualquer tipo de dado de qualquer tamanho pode ser copiado em qualquer taxa de dados usando qualquer método de importação (inicial, lote, *streaming*) em seu formato original (nativo, cru).

Data Lake: Teoria

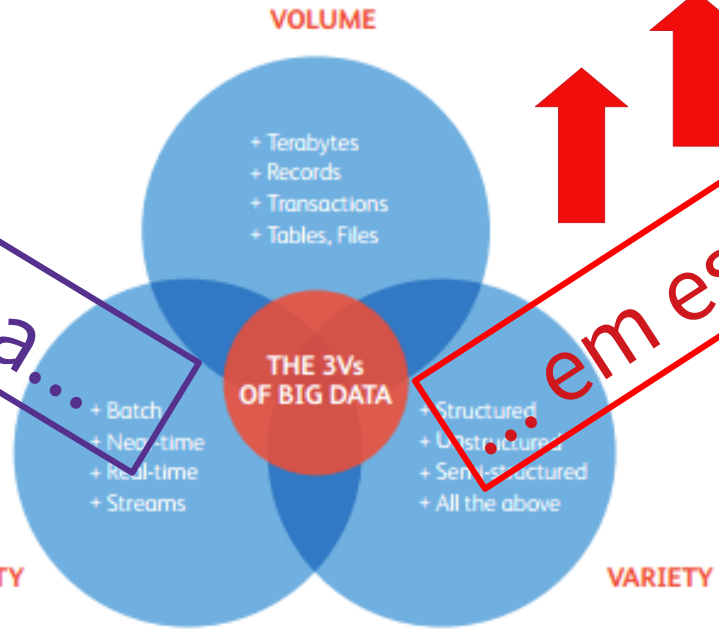
THE DEVELOPER'S CONFERENCE

- Então... o que é exatamente um *Data Lake*?
  - O *Data Lake* como arquitetura refere-se a um sistema ou ecossistema (Elliott, 2015; Intersog, 2016; Khanna, 2016; O'Brien, 2015; Rivera, 2014; TeraData, 2014)
  - Uma combinação de múltiplos sistemas que se integram intimamente e servem ao mesmo propósito (TeraData, 2014)
  - *Data Warehouse* para Big Data

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

Mesmo problema...



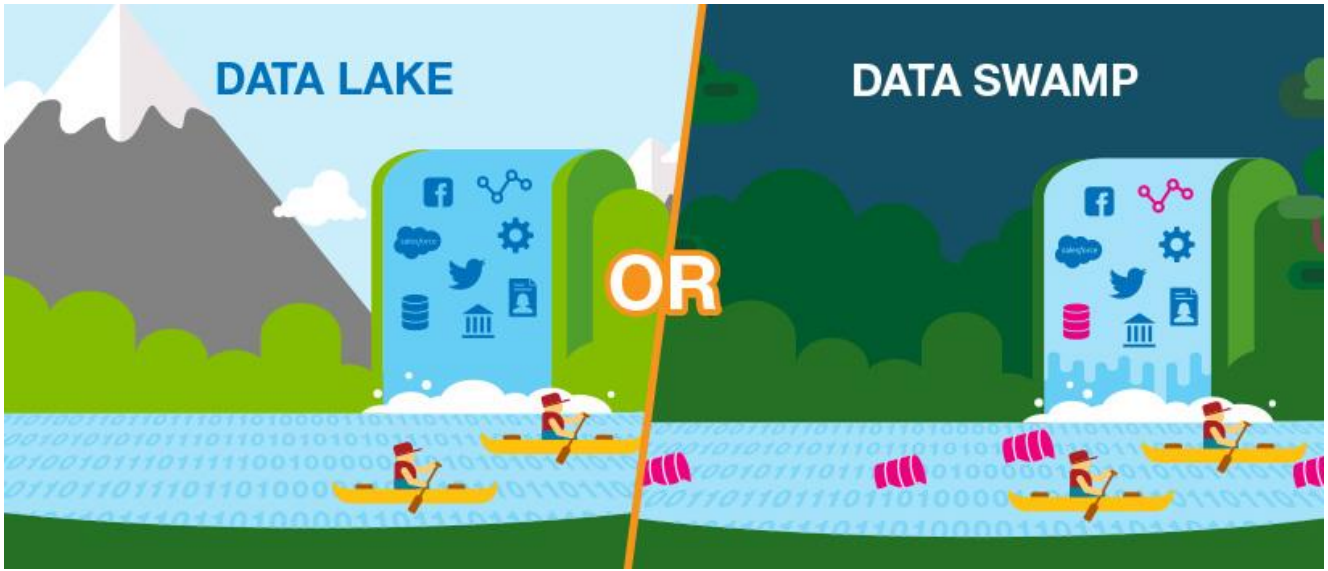
...em escala MAIOR

E os Vs?

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

Data Swamps? Como assim?



Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

- Ok, eu entendi... mas de maneira menos figurativa!
  - **É descrito como um "armazenamento de dados despejados", aonde "despejo" significa que os dados são fornecidos sem contexto, metadados nem algum tipo de controle sobre os dados.**
  - **Os *Data Lakes* possuem as maiores chances de se tornarem parcial ou totalmente *Data Swamps*!**
  - **Data Swamps tornam os dados inúteis, porque o dado não pode ser analisado (Intersog, 2016)**

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

- Inúteis? Mas como dados podem ser inúteis?
  - sem senso de governança de dados
  - sem a devida garantia de qualidade de metadados e dados

Sem organização...

... não há conhecimento

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

- Então... Precisamos evitar pântanos!
- *Data Warehouses* então não são melhores?

> *Data Lakes* → *Big Data*
>
> *Data Warehouse* → **Bancos de Dados Relacionais**

- *Data Warehouses* e *Data Lakes* são conceitos completamente diferentes!

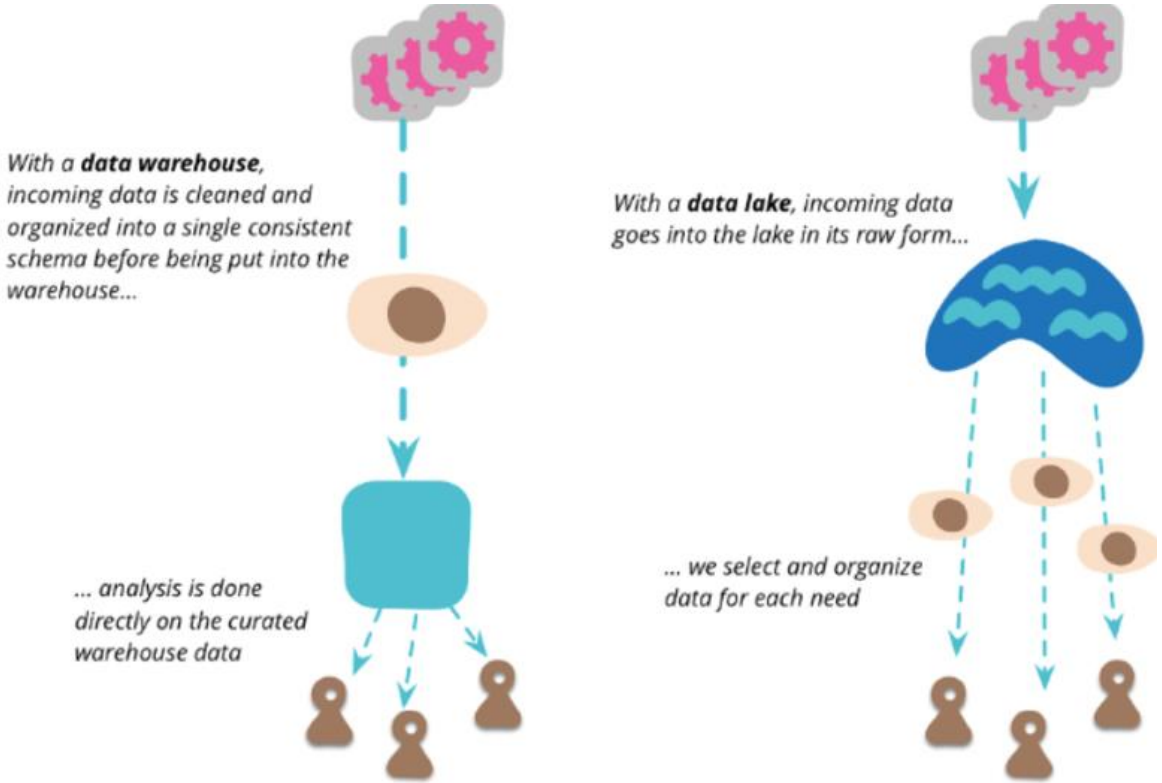
Data Lake: Teoria

THE DEVELOPER'S CONFERENCE

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Data Lake: Teoria

THE DEVELOPER'S CONFERENCE



Data Lake: Prática

THE DEVELOPER'S CONFERENCE

Que produtos nós temos envolvendo *Data Lakes*, atualmente?

Microsoft

amazon

ORACLE®

Data Lake: Prática

THE DEVELOPER'S CONFERENCE

Data Lake: Prática

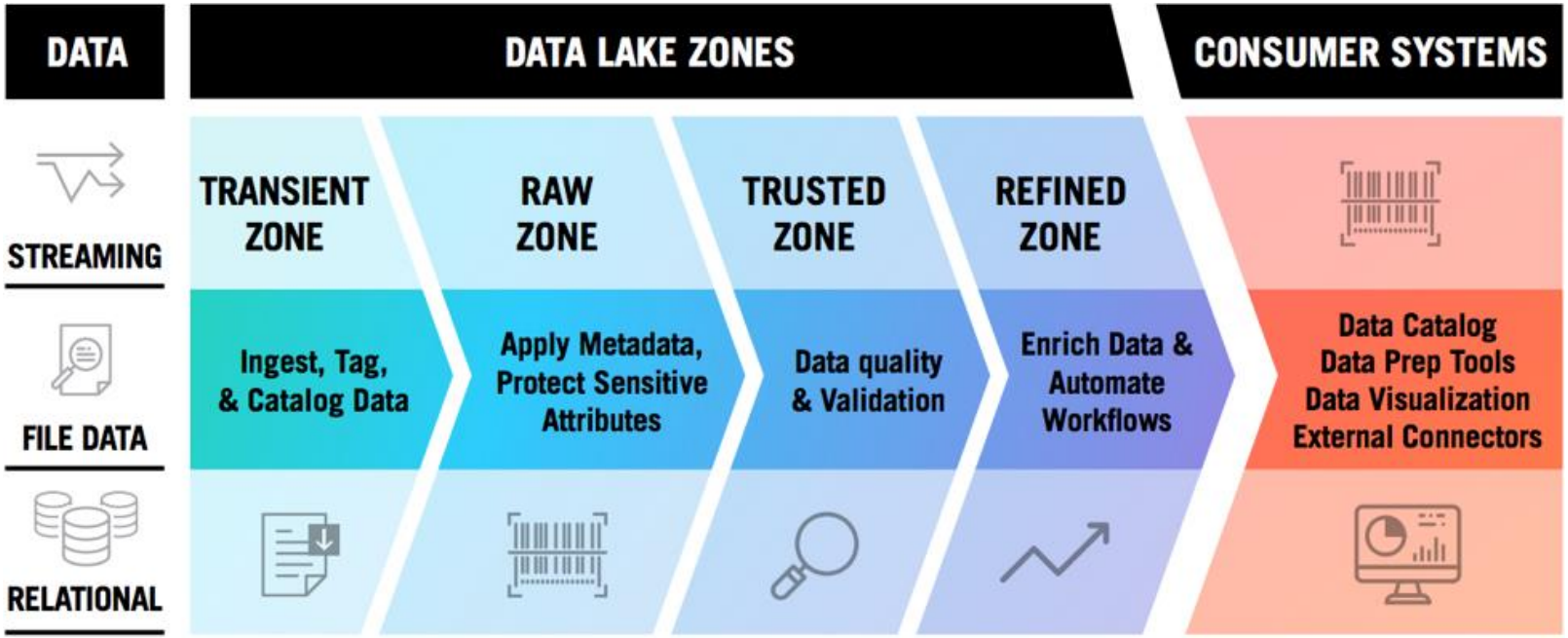
THE DEVELOPER'S CONFERENCE

- Maioria das implementações são baseadas em *Hadoop* apenas!
- Muitas delas não conseguem lidar com os vários tipos de dados!
- Conceitualmente falando, usa uma grande variedade de tecnologias (e modelos de bancos de dados) para alcançar seu objetivo!

Data Lake: Prática

THE DEVELOPER'S CONFERENCE

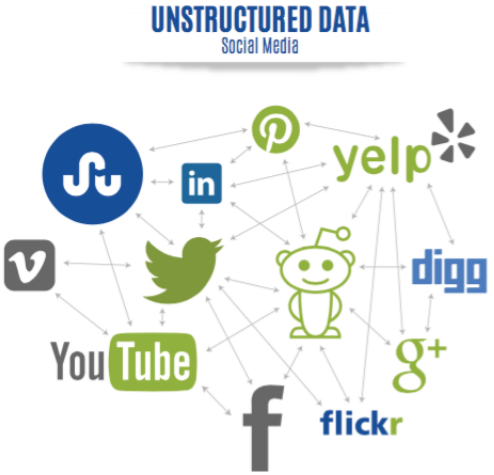
Ok, o que um *Data Lake* eficiente deve fazer?



Data Lake: Prática

THE DEVELOPER'S CONFERENCE

- O que esse "*Data Lake* conceitual" pode fazer no mundo real?
  - **Bancos de Dados Multi-Modelo**
  - **Lidar com quaisquer tipos de dados**
  - ***Machine Learning***



Data Lake: Prática

THE DEVELOPER'S CONFERENCE

- Exemplos, por favor!
  - **Banco de Dados Relacionais, integrados com o *Lake* (MySQL, PostgreSQL, Oracle, SQL Server, DB2...)**
  - **Bancos de Dados NoSQL, também integrados ao *Lake* (Cassandra, MongoDB, Neo4J…)**
  - **Frameworks de Processamento de Dados lidando com os dados (Hadoop, Spark, Flink, Storm…)**
  - **Algoritmos de Machine Learning, preferencialmente Computação BioInspirada (Redes Neurais, Algoritmos Genéticos, Inteligência de Enxame...)**

Data Lake: Prática

THE DEVELOPER'S CONFERENCE

- Então, não há tecnologia específica
  - Precisamente. O *Data Lake* é conceitual, e vem de uma necessidade de processar e lidar com grandes quantidades de dados
  - Muitas das tecnologias são, de fato, livres! Baseado nesse conceito, um *Lake* pode ser criado sem qualquer software mandatório que precise de licença paga
  - Claro, de acordo com sua variedade, volume e velocidade de dados (3V's), você irá precisar investir em *hardware*

Referências

- Dixon J. (2010) **Pentaho, Hadoop, and Data Lakes**. Lido em: 20 de Março de 2019. De: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
- Elliott, T. (2015). **From Data Lakes to Data Swamps**. Lido em: 21 de Março de 2019. De: http://www.zdnet.com/article/from-data-lakes-to-data-swamps
- Intersog (2016). "**What Is The Difference Between Data Lakes, Data Marts, Data Swamps, And Data Cubes?"** Lido em: 19 de Março de 2019. De: http://intersog.com/blog/what-is-the-difference-between-data-lakes-data-marts-data-swamps-and-data-cubes/

Referências

- Khanna, A. (2016, April 15). **How to Keep Your Data Lake From Becoming a Data Swamp**. Lido em: 18 de Março de 2019. http://www.reltio.com/about/news/2016/4/how-to-keep-your-data-lake-from-becoming-a-data-swamp
- Rivera, J. (2014, July 28). **Gartner Says Beware of the Data Lake Fallacy**. Lido em: 18 de Março de 2019. https://www.gartner.com/en/newsroom/press-releases/2014-07-28-gartner-says-beware-of-the-data-lake-fallacy
- TeraData, HortonWorks. (2014). **Putting the Data Lake to Work**: a Guide to Best Practices. Lido em: 20 de Março de 2019. https://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf

THE DEVELOPER'S CONFERENCE

Jorge Sandoval
M.Sc. em Computação Aplicada