

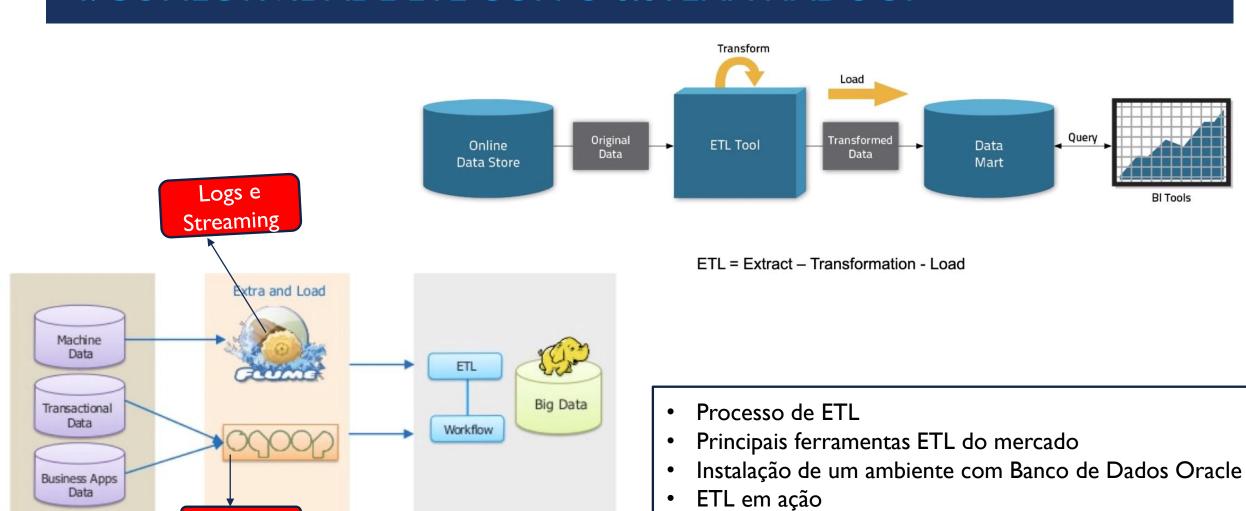
ENG DE DADOS COM HADOOP E SPARK 4

1. Planejando e Configurando um Cluster Hadoop 2. Usando MapReduce em Grandes Volumes de Dados 3. Armazenamento de dados com HBase e Hive 5. Administração e Manutenção do Hadoop 6. Hadoop Machine Learning com Apache Mahout 7. Apache Hadoop e Apache Spark



4. CONECTIVIDADE ETL COM O SISTEMA HADOOP

Batch

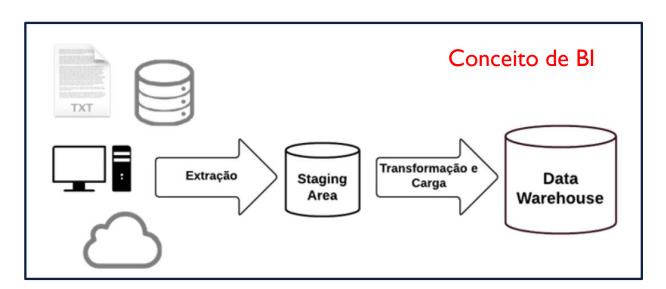


BI Tools

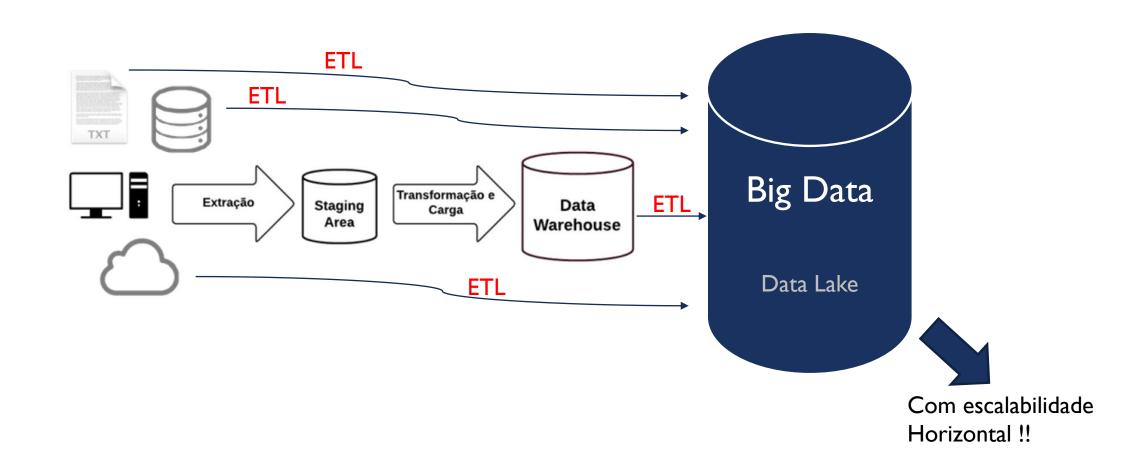
4.1 O QUE É ETL?

• ETL fornece a infraestrutura de integração através da realização de três importantes funções:





4.2 QUAL O PAPEL DO ETL NO BIG DATA?



4.3 PRINCIPAIS FERRAMENTAS ETL DO MERCADO

ETL é atribuição do Engenheiro de Dados!

Principais Ferramentas ETL - Proprietárias

- Informatica Power Center
- IBM InfoSphere Data Stage
- Oracle Data Integrator (ODI)
- Microsoft SQL Server Integration Services (SSIS)
- SAS Data Integration Studio
- SAP Business Object Integrator
- Pentaho Data Integration

Principais Ferramentas ETL - Open Source

- Dataiku Data Science Studio (DSS) Community Edition
- Talend Open Studio For Data Integration
- Jaspersoft ETL
- Jedox
- RapidMiner
- Apache Flume
- Apache NiFi
- Apache Sqoop

ETL em tempo real!

Principal ferramenta ETL do ecossistema Hadoop para carga de dados em Batch (lote)

4.4 PRINCIPAIS BANCO DE DADOS DO MERCADO

364 systems in ranking, March 2021

Rank				Score		
Feb 2021	Mar 2020	DBMS	Database Model	Mar 2021	Feb 2021	Mar 2020
1.	1.	Oracle 😷	Relational, Multi-model 🔃	1321.73	+5.06	-18.91
2.	2.	MySQL [Relational, Multi-model 👔	1254.83	+11.46	-4.90
3.	3.	Microsoft SQL Server [1]	Relational, Multi-model 🔞	1015.30	-7.63	-82.55
4.	4.	PostgreSQL [+	Relational, Multi-model 👔	549.29	-1.67	+35.37
5.	5.	MongoDB 🚹	Document, Multi-model 📵	462.39	+3.44	+24.78
6.	6.	IBM Db2 ₽	Relational, Multi-model 📵	156.01	-1.60	-6.55
7.	1 8.	Redis 🚻	Key-value, Multi-model 🔞	154.15	+1.58	+6.57
8.	4 7.	Elasticsearch 😷	Search engine, Multi-model 👔	152.34	+1.34	+3.17
9.	1 0.	SQLite [Relational	122.64	-0.53	+0.69
↑ 11.	4 9.	Microsoft Access	Relational	118.14	+3.97	-7.00
↓ 10.	11.	Cassandra 🚹	Wide column	113.63	-0.99	-7.32
12.	1 3.	MariaDB 🚹	Relational, Multi-model 🔞	94.45	+0.56	+6.10
13.	4 12.	Splunk	Search engine	86.93	-1.61	-1.59
14.	14.	Hive	Relational	76.04	+3.72	-9.34
1 6.	15.	Teradata	Relational, Multi-model 🔞	71.43	+0.53	-6.41
4 15.	↑ 23.	Microsoft Azure SQL Database	Relational, Multi-model 🔞	70.88	-0.41	+35.44
	Feb 2021 1. 2. 3. 4. 5. 6. 7. 8. 9. ↑11. ↓10. 12. 13. 14. ↑16.	Feb 2021 Mar 2020 1. 1. 2. 2. 3. 3. 4. 4. 5. 5. 6. 6. 7. ↑ 8. 8. ↓ 7. 9. ↑ 10. ↑ 11. ↓ 9. ↓ 10. 11. 12. ↑ 13. 13. ↓ 12. 14. 14. ↑ 16. 15.	Feb 2021 Mar 2020 DBMS 1. 1. Oracle : 2. 2. MySQL : 3. 3. Microsoft SQL Server : 4. 4. PostgreSQL : 5. 5. MongoDB : 6. 6. IBM Db2 : 7. ↑ 8. Redis : 8. ↓ 7. Elasticsearch : 9. ↑ 10. SQLite : 11. ↓ 9. Microsoft Access ↓ 10. 11. Cassandra : 12. ↑ 13. MariaDB : 13. ↓ 12. Splunk 14. 14. Hive ↑ 16. 15. Teradata	Feb 2021Mar 2020DBMSDatabase Model1.1.Oracle ↑Relational, Multi-model ↑2.2.MySQL ↑Relational, Multi-model ↑3.3.Microsoft SQL Server ↑Relational, Multi-model ↑4.4.PostgreSQL ↑Relational, Multi-model ↑5.5.MongoDB ↑Document, Multi-model ↑6.6.IBM Db2 ↑Relational, Multi-model ↑7.↑ 8.Redis ↑Key-value, Multi-model ↑8.↓ 7.Elasticsearch ↑Search engine, Multi-model ↑9.↑ 10.SQLite ↑Relational11.↓ 9.Microsoft AccessRelational↓ 10.11.Cassandra ↑Wide column12.↑ 13.MariaDB ↑Relational, Multi-model ↑13.↓ 12.SplunkSearch engine14.14.HiveRelational↑ 16.15.TeradataRelational, Multi-model ↑	Feb 2021 Mar 2021 1. 1. Oracle ↑ Relational, Multi-model ↑ 1321.73 2. 2. MySQL ↑ Relational, Multi-model ↑ 1254.83 3. 3. Microsoft SQL Server ↑ Relational, Multi-model ↑ 1015.30 4. 4. PostgreSQL ↑ Relational, Multi-model ↑ 549.29 5. 5. MongoDB ↑ Document, Multi-model ↑ 462.39 6. 6. IBM Db2 ↑ Relational, Multi-model ↑ 156.01 7. ↑ 8. Redis ↑ Key-value, Multi-model ↑ 154.15 8. ↓ 7. Elasticsearch ↑ Search engine, Multi-model ↑ 152.34 9. ↑ 10. SQLite ↑ Relational 122.64 ↑ 11. ↓ 9. Microsoft Access Relational 118.14 ↓ 10. 11. Cassandra ↑ Wide column 113.63 12. ↑ 13. MariaDB ↑ Relational, Multi-model ↑ 94.45 13. ↓ 12. Splunk Search engine 86.93	Feb 2021 Mar 2020 DBMS Database Model Mar 2021 Feb 2021 1. 1. Oracle ↑ Relational, Multi-model ↑ 1321.73 +5.06 2. 2. MySQL ↑ Relational, Multi-model ↑ 1254.83 +11.46 3. 3. Microsoft SQL Server ↑ Relational, Multi-model ↑ 1015.30 -7.63 4. 4. PostgreSQL ↑ Relational, Multi-model ↑ 549.29 -1.67 5. 5. MongoDB ↑ Document, Multi-model ↑ 462.39 +3.44 6. 6. IBM Db2 ↑ Relational, Multi-model ↑ 156.01 -1.60 7. 8. Redis ↑ Key-value, Multi-model ↑ 154.15 +1.58 8. √ 7. Elasticsearch ↑ Search engine, Multi-model ↑ 152.34 +1.34 9. ↑ 10. SQLite ↑ Relational 122.64 -0.53 ↑ 11. ✓ 9. Microsoft Access Relational, Multi-model ↑ 13.45 +0.56 13. ✓ 12. Splunk

4.5 OPERAÇÃO DE ETL COMPLETA COM BANCO DE DADOS ORACLE E APACHE SQOOP

O intuito desse projeto é montar um banco de dados relacional, simulando um data Warehouse com banco de dados Oracle para na sequência carregar 20 milhões de registros nesse banco de dados e então utilizar o Apache Sqoop como ferramenta ETL para levar os dados do Banco Oracle para o HDFS.





THANKS