简介

- snowflake 和 databrick 两款产品资料收集。
- 产品选型上，两款产品面向的用户群体是不一样的，存在较为明显的区别，相似用户群在数据分析师。
- 在技术选型上，两款产品的技术路线存在不同，导致彼此之间存在优缺点之间的互补。
- 从用户反馈上看，snowflake的评分要高于databrick。

目录

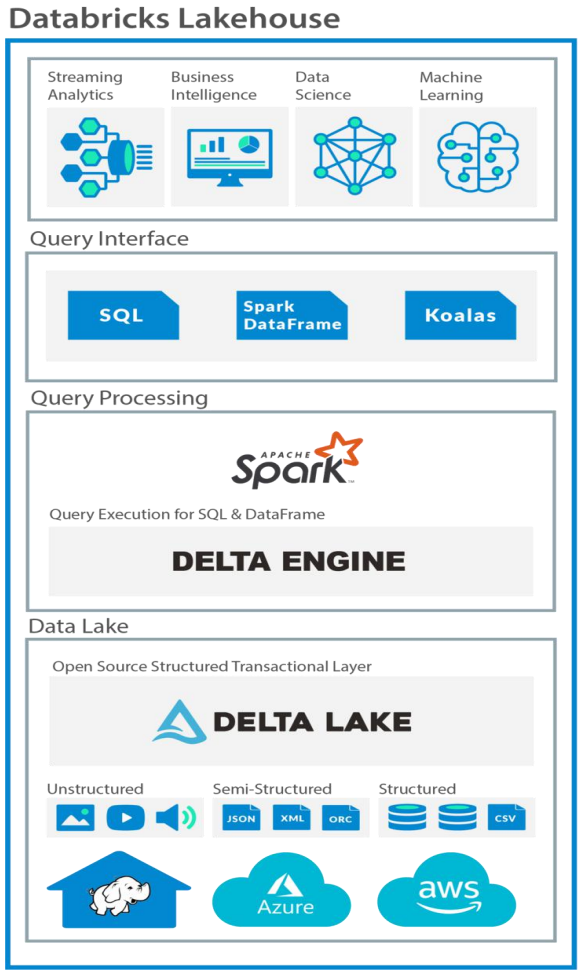
1.架构-Snowflake vs DataBricks

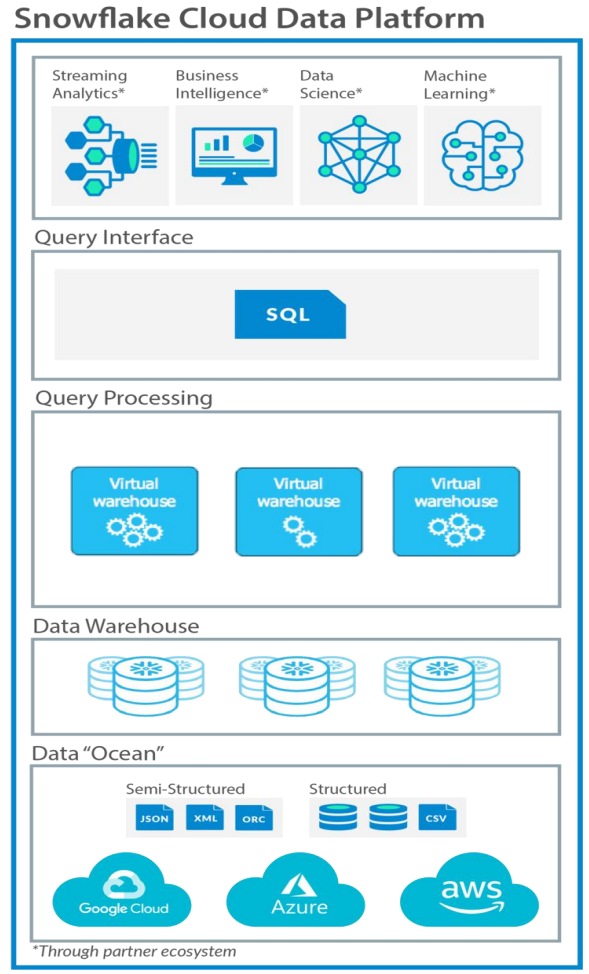
snowflake:

- 面向数据仓库
  结构化、半结构数据等
- 主要面向数据开发人员
- mpp
- 抽象化帮助理解:
  将架构抽象成存储、计算和服务。
用户只需关心服务层面，而不用下沉到具体的计算资源、存储资源等层面
- 安全:
  端到端的计算链、数据存储
- 公有云为主
- 以sql为主
- 按需计费、事后计费

databrick:

- 面向数据湖
  结构化、半结构化、无结构数据等
- 主要面向数据分析和AI工程师，主打data+AI
- spark
- 用户不仅需要关心服务层面，也需要关心一些底层云设施
- 支持混合云
- sql\python等多语言





1.架构-参考

- snowflake与databrick对比:https://www.datagrom.com/data-science-machine-learning-ai-blog/snowflake-vs-databricks
- snowflake论文:http://pages.cs.wisc.edu/~yxy/cs839-s20/papers/snowflake.pdf
- 论文解读和引申:https://zhuanlan.zhihu.com/p/56745552

2.产品的形态

- SnowFlake 简洁、易上手，侧重于数仓，适用于数据工程师和分析师等。
  - introduction：https://resources.snowflake.com/youtube-all-videos/snowflake-introduction-demo
  - demo：https://resources.snowflake.com/youtube-all-videos/concurrency-in-snowflake

- DataBrick 简洁，功能较多，侧重于分析，适用于数据科学家、数据分析师等。
  - sql： https://databricks.com/discover/demos/databricks-sql
  - demos：https://databricks.com/discover/demos

3.用户场景 - snowflake

- 小中大型数据收集公司：

  数据共享；跨地域数据获取；

- 商业智能团队：

  营销与产品决策等数据的获取与简要分析

- 中小型数仓团队

3.用户反馈-snowflake

技术优点

- 云：扩展性、计算性能、免维护，易于云端集成，存储和计算分离。
- 数据的存储、组织、管理方便；与aws s3 无缝集成；
- 近实时的查询，优化的查询引擎、支持标准的ANSI sql语言
- 即使账号内的员工数增加、也不影响系统的稳定性。
- 数据迁移方便、易与其他数据中心连接；
- 无索引的设计，可以让用户不用花时间在管理和维护解决方案上，而是关注业务。

产品优点：

- 安全的数据共享： 与供应商间的数据传输，与公司内部经理间数据共享（工作协同）；与业务方共享业务看板/商业洞察；简单易用且透明；
- 易于与其他bi分析工具的联动等（tableau\power bi\sql server）；还可以将用户的sql ide连接到snowflake
- 友好的用户查询界面，上手简单、易于初学者。查询建议、表可以拖拽到查询中。
- 免费试用
- 成本控制：按需计费；计算和存储计费分开；

3.用户反馈-snowflake

- **缺点**

  - 不能恢复误删的查询表
  - 文档不完善
  - 不能支持PLSQL
  - 图表功能有待改进、报表功能需要跟第三方联动；
  - 可备选的产品：Informatica Enterprise Data Integration 、TreasureData
  - 不适用于没上云的企业（感觉是公有云的锅）
  - 没有存储查询代码的功能，需要重新查询
  - 多用户同时使用，存在性能瓶颈（感觉是mpp的锅）
  - 技术支持响应不及时
  - 不支持Dynamic SQL
  - 不支持python等语言
  - 复杂查询速度缓慢
  - 用户业务的SLA会受三大云商的影响
  - 任务调度功能不足

4.用户场景-databrick

- 大型的数据转换与分析：同时处理海量数据、可视化
- 数据科学家团队

4.用户反馈-databrick

- 优点:
  - 与Azure无缝集成
  - 数据加载、提取、转换、集成的处理速度快,
  - 多语言支持（python等）
  - 界面简单
  - github的集成
  - DAG任务调度管理
  - ML flow
  - 灵活的编写代码和自动化运行任务
  - 帮助托管spark集群
  - notebook对协同工作有帮助，类似terminal便于编码环境。
  - 数据处理到模型训练无缝集成

4.用户反馈 - databrick

- 缺点:
  - 没有免费试用环节
  - 背后强依赖scala语言
  - 价格相对高
  - 单元测试支持不好
  - 无法直接更新云数据库中的数据
  - 集群创建时间漫长
  - 存在非按需付费的问题。
  - 不容易理解（学习成本太高）
  - 数据分析或挖掘时，参数还是手动，希望可以自动
  - 错误提示不明显，往往需要进一步去查找

4.场景与用户反馈资料参考

- sourceforge： https://sourceforge.net/software/compare/Cloudera-vs-Databricks-vs-Snowflake/
- g2： https://www.g2.com/compare/azure-databricks-vs-snowflake
- trustradius:https://www.trustradius.com/compare-products/databricks-lakehouse-platform-vs-snowflake
- etl角度考虑:https://www.confessionsofadataguy.com/databricks-vs-snowflake-the-datalake-warehouse-battle/
- kylingence关于snowflake的分析： https://kyligence.io/blog/snowflake-the-good-the-bad-and-the-beautiful-for-analytics/