AI Pioneer

2018 AI Pioneer Conference

Introduction to Huawei Cloud Data Lake

Wang Fei

Agenda

- Data Lake Introduction
- Data Lake Architecture on Huawei Cloud
- Data Lake Insight introduction
- Huawei Case Study

What is a Data Lake

- Data mart/data warehouse

  A store of bottled water – cleansed, packaged and structured for easy consumption.

What is a Data Lake

- Data Lake

  A large body of water in a more natural state.

What is a Data Lake

- A data lake is a single place to put all the data enterprises want to collect, store, analyze and turn into business insights and decisions, including structured, semi-structured and unstructured data.

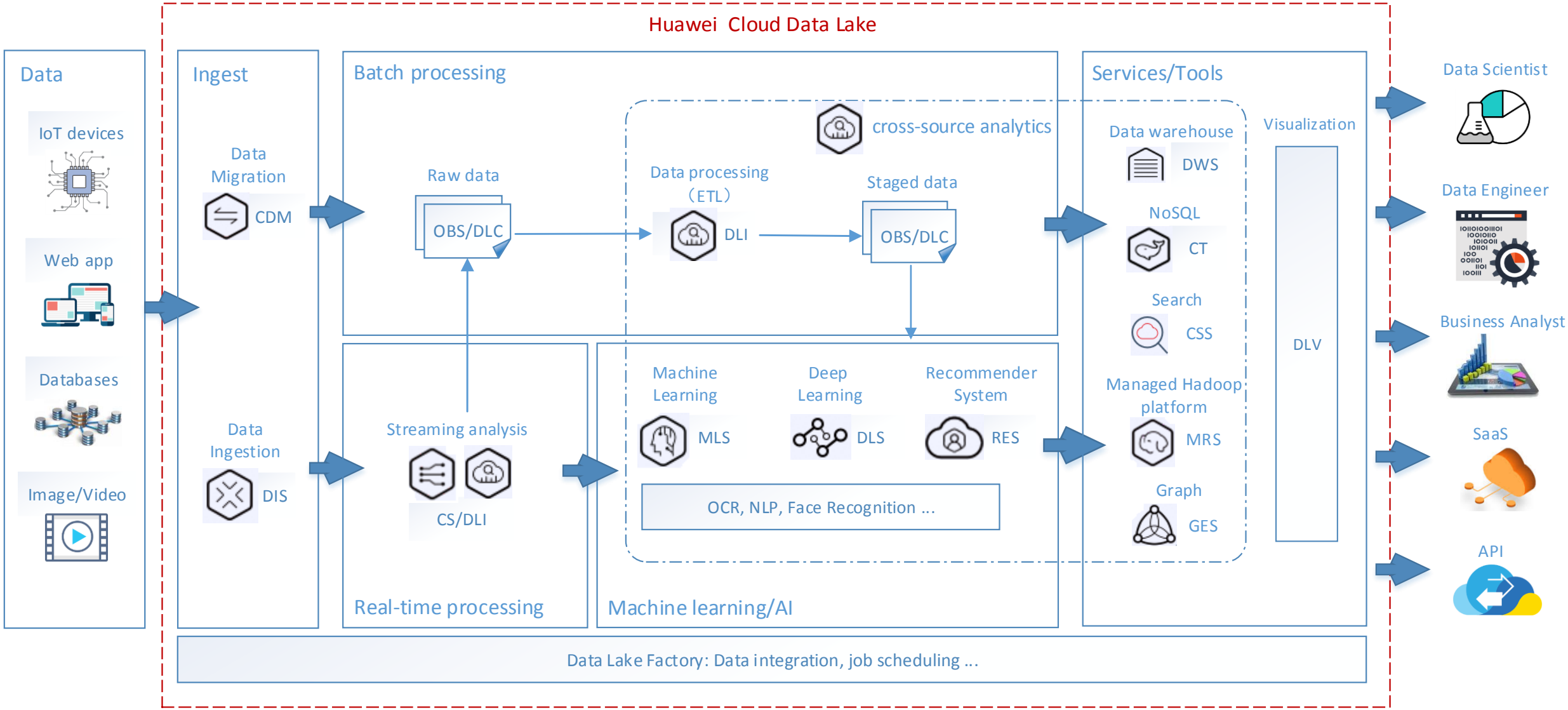
Why Data Lake

- **Increasing data volumes** – provide a scalable, cost-effective solution to gather, store and analyze data.
- **Various data sources and formats** – deal with structured, semi-structured and unstructured data by utilizing the most appropriate tools for different purposes.
- **Time from data to business value** – rapid ingestion and processing of data, e.g. real-time analysis.
- **Data collaboration** – model and publish governed data sets for easy visualization and access.

Agenda

- Data Lake Introduction
- Data Lake Architecture on Huawei Cloud
- Data Lake Insight introduction
- Huawei Case Study

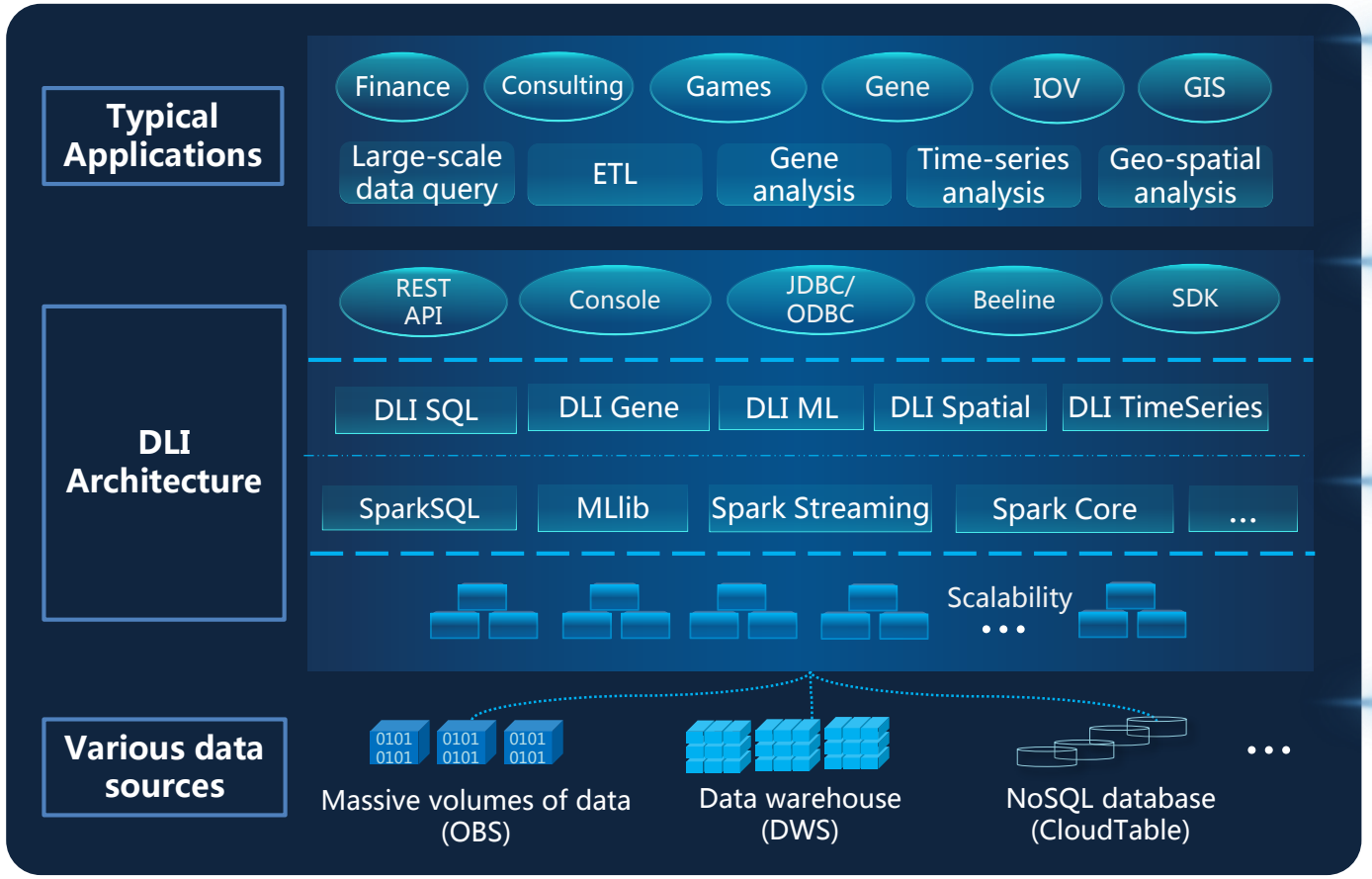
Huawei Cloud Data Lake Architecture



Agenda

- Data Lake Introduction
- Data Lake Architecture on Huawei Cloud
- Data Lake Insight introduction
- Huawei Case Study

DLI: a fully managed, highly scalable and unified data analytics platform



Fully Managed and Ease of Use

Users can enjoy data analytics on DLI with zero maintenance cost and no hardware/software management headache.

Fully Compatible with Big Data Ecosystem

Provides full-stack Apache Spark capabilities and APIs. Users can perform SQL, machine learning, streaming applications on one unified platform.

Federated Data Analytics

Supports cloud-based federated data analytics directly with data residing on local storage, OBS, DWS, CloudTable, etc.

Best Practice Analysis Toolkits

Provides best practice analysis toolkits and standard APIs in many industry domains including Genomics, Geographics, etc.

DLI Summary

- DLI has the philosophy to minimize complexity and simplify user experience: fully managed and serverless.
- DLI's focus is on analytic processing. DLI is SQL2003 compliant, and goes beyond SQL with built-in machine learning and streaming support.
- DLI is fully compatible with Apache Spark ecosystem, allowing users to perform existing Spark workloads. However, DLI is different from Apache Spark in following aspects:
  - Multi-tenant management: role management and access control.
  - SQL optimizations: various new optimization rules, cost-based optimization (contributed to Apache Spark).
  - Enhanced data sources: data warehouses, relational databases, unstructured data, etc.
  - Integrated analysis toolkits: GATK, GeoMesa, OpenTSDB, etc.

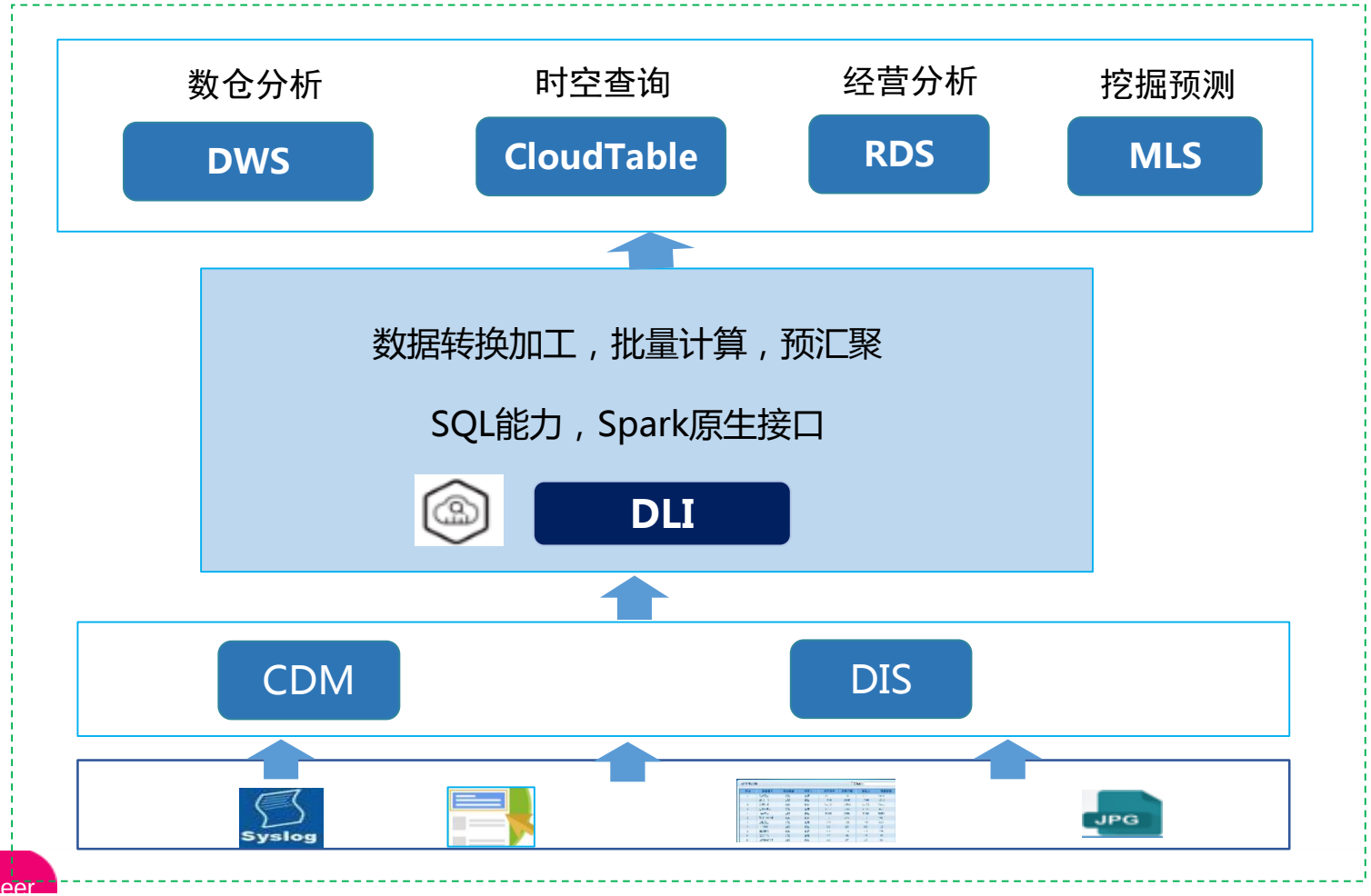
Agenda

- Data Lake Introduction
- Data Lake Architecture on Huawei Cloud
- Data Lake Insight introduction
- Huawei Case Study

大数据ETL处理

应用场景简介：

随着信息化时代的来临，企业数据产生越来约迅速，数据种类越来越多，体量也越来越大，如何挖掘数据的价值是企业信息化首要问题，在数据变成价值之前，往往需要对大量数据进行加工转换，预先汇聚等各种ETL处理



方案特点：

- 多类型数据支持：结构化/半结构化/非结构化数据计算处理
- 海量数据计算处理：计算弹性扩展，数据规模从GB~EB不等，真正的大数据处理
- 业务处理灵活：同时提供SQL和编程的方式
- 生态开放：完全兼容开源Spark社区原生接口，业务迁移平滑
- 按需计费：真正的按需计算，使用时收费，不使用不收费；有效应对业务波峰波谷的成本控制

多数据源融合探索分析

应用场景简介：

随着信息化时代的来临，企业数据产生越来约迅速，数据种类越来越多，体量也越来越大，如何挖掘数据的价值是企业信息化首要问题，在数据变成价值之前，往往需要对大量数据进行加工转换，预先汇聚等各种ETL处理

方案特点：

- 统一分析入口：多种应用统一分析入口，数据统一逻辑视图
- 多数据源，免搬迁：数据格式多样，存储在各个系统中，直接访问，不用搬迁即可进行分析
- AI能力低门槛使用：内部关联AI能力，如文字识别，图像识别等，用户通过SQL即可使用AI，进行非结构化数据分析
- 企业多租户：资源按租户隔离，数据权限控制到表、列，方便用户进行精细化管理

线下Spark/Hive迁移场景

客户群：自建数据中心Spark/Hive企业，数据平台公司，IT人员建设不多
方案特点：接口完全兼容，业务搬迁零改动，超低成本，免运维，功能比社区还强大

自建机房

自建DC，高额水电消耗

人工维护

安装部署3天
集群升级扩容超过一周

技术门槛高

100万+行代码，数百参数，专业调优
社区版本升级不兼容，业务需改动

成本浪费

固定集群，业务低峰造成浪费

线下自建
Spark集群

VS

华为云
DLI服务

华为8年技术积累
社区贡献全球TOP4
Spark核心优化器Commiter
Gartner魔力象限，中国TOP1

- DLI计算性能是Hive2~5倍，让您享受极速体验
- DLI在社区Spark基础上进行稳定可靠，安全上完成企业化商用改造，让你用得放心

不用申请机器构建机房
用户即来即用，用完即走

无需建机房

不Care软件升级扩容
运维交给专业的华为云

免运维

无需关注技术细节，直接调接口
100%兼容社区，业务“0”改动迁移

低门槛

按需使用，扫描1GB3分钱
用多少付费多少，不用不收钱

超低成本

Questions & Discussion

THANKS