生态与兼容

MaxCompute大数据生态集成和开发工具

薛明/艺卓
阿里云数据事业部高级专家

当我们在说生态和兼容时

1

上云工具

OGG、Sqoop、Flume、FluentD

2

社区兼容

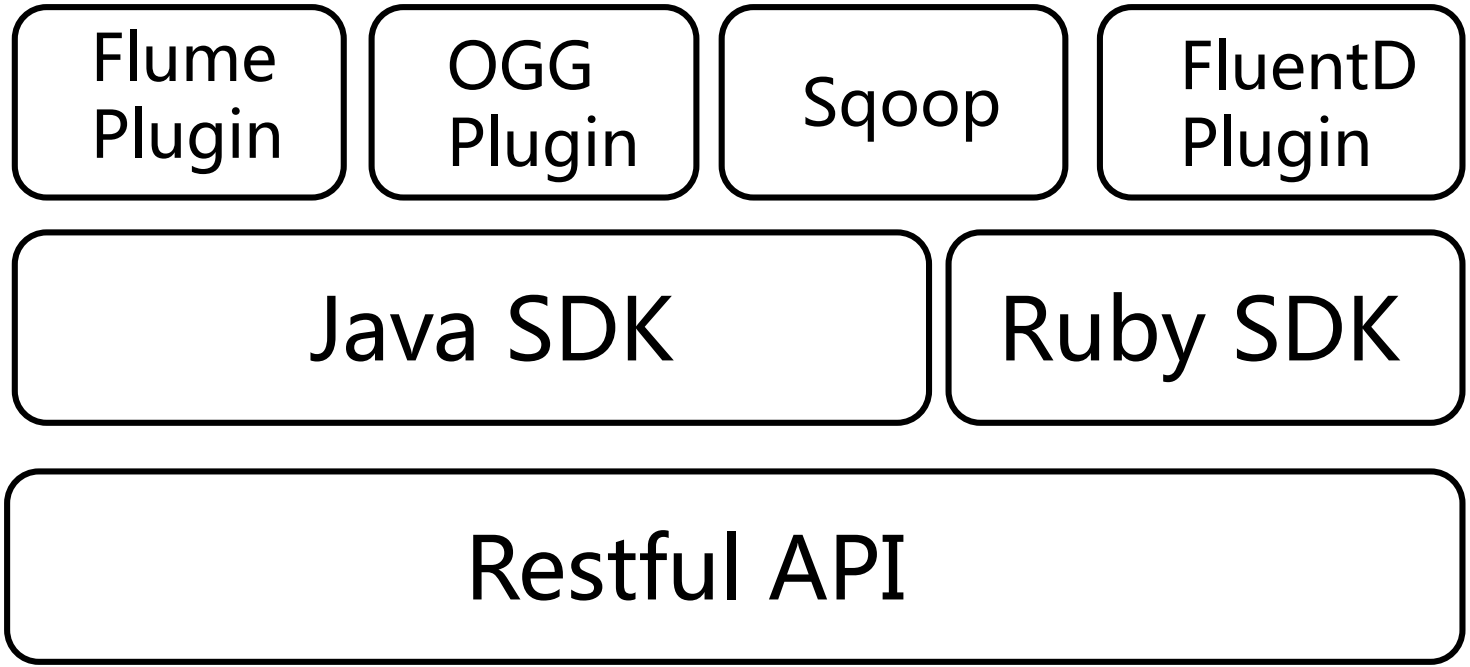
SQL、Hadoop MR、Hive Thrift 协议

3

生态连接

JDBC、ODBC、R、Python Pandas、IntelliJ IDEA

依托开源社区的上云工具



社区兼容

1

SQL

SQL 语法，UDF，文件格式

2

Hadoop MR Adapter

只需替换一个 JAR 包，
运行时翻译为 MaxCompute OpenMR 执行

3

HiveProxy

提供 Hive Thrift 协议兼容层，
可以直接对接 Beeline、Hive ODBC+Tableau

生态连接能力

1

协议

JDBC、ODBC

2

语言

R、Python Pandas

3

集成开发环境

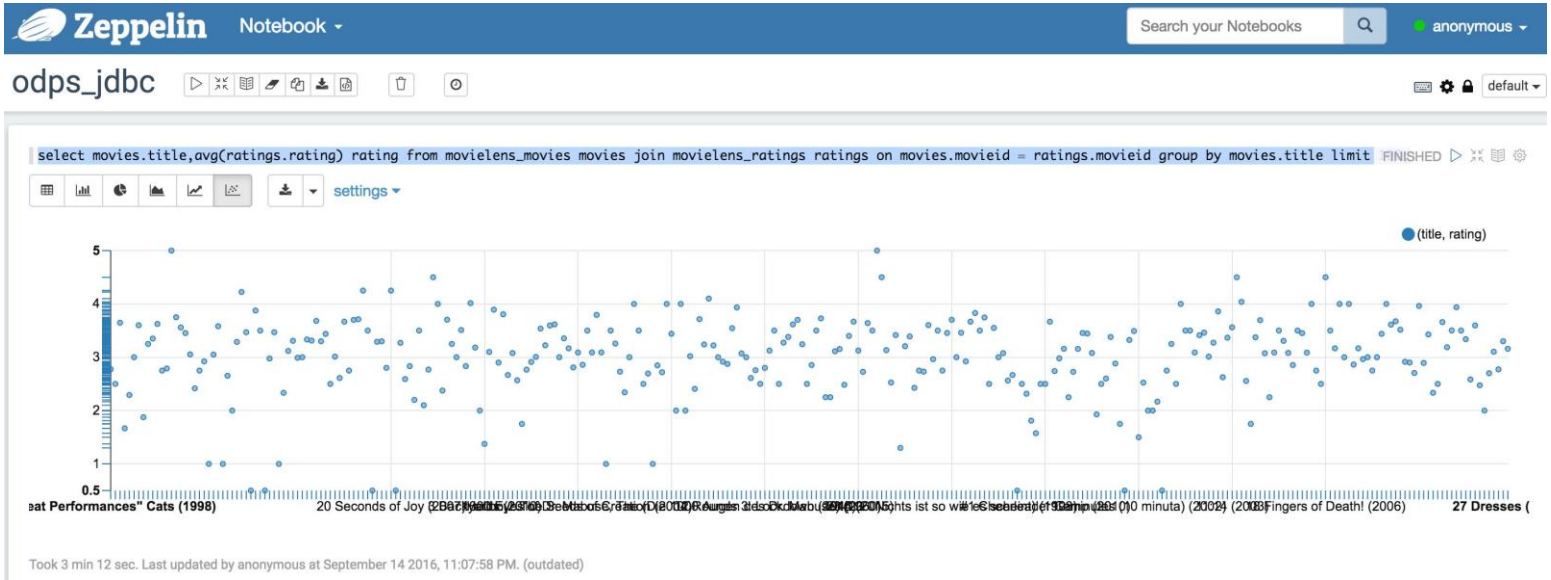
IntelliJ IDEA 插件

连接生态能力：JDBC

使支持 JDBC 的工具可以直接支持 MaxCompute：如 BI 工具 Pentaho；数据库管理工具 SQL Workbench；交互式数据分析工具 Zeppelin，等等。

使用标准 JDBC SDK 编程实现简单的数据查询任务。

连接生态能力：JDBC



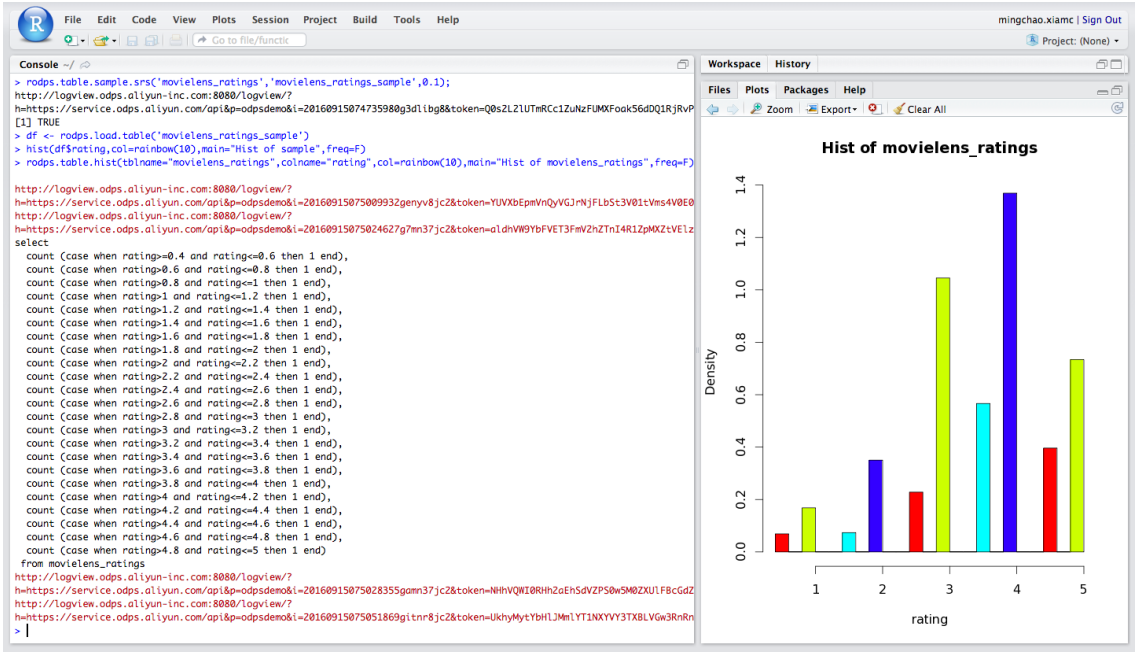
扫码观看大会视频

连接生态能力：RODPS

RODPS 使 R 用户具备了直接执行 MaxCompute SQL，并将结果集转换为 R data.frame 的能力。

因此，RODPS 很适用于先在 MaxCompute 中完成大数据量的过滤等操作，缩小数据规模，再使用 R 来进行单机分析的场景。

连接生态能力：RODPS



扫码观看大会视频

连接生态能力：PyODPS

PyODPS 不仅仅是 MaxCompute 的 Python SDK，它还为熟悉 Pandas 的用户提供了基于 MaxCompute SQL 的分布式 DataFrame 的执行能力。

PyODPS 提供了对 PAI 算法的集成，可以用简单的几行代码来实现复杂的算法调用流程。

PyODPS 还提供了基于 Jupyter Notebook 的交互式数据分析能力。

连接生态能力：PyODPS

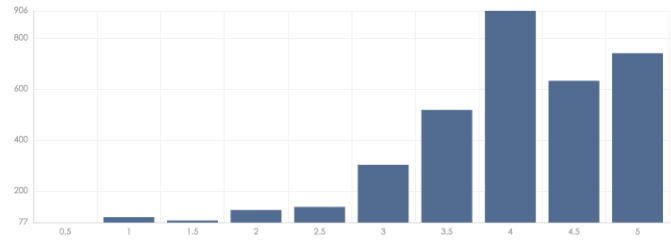
分析标签数据

```
In [9]: df = ratings.join(tags, on=['userid', 'movieid'], suffixes=('', '1'))
review = df.groupby(ratings.dtypes.names).agg(
    tags=(df.tag + ',').sum().slice(0, -1),
    timestamps=(df.timestamp.astype('string') + ',').sum().slice(0, -1)
)
review.head(10)
```

userid	movieid	rating	timestamp	tags	timestamps
159	96481	2	1421443862	white trash,documentary,lifestyle	1421443862,1421443862,1421443862
178	70	3	1165634696	Diablo	1165634696
178	39292	3	1140216203	borring?	1140216203
178	49286	3	1169680582	Jack Black ^_^	1169680582
206	1260	4.5	1287251644	Fritz Lang	1287251644
206	2318	4	1287247624	dysfunctional family,dark comedy,dark humor	1287247624,1287247624,1287247624
261	1	4.5	1313154689	animation,fun	1313154689,1313154689
261	356	4	1313154447	drama	1313154447
261	1222	3	1313150742	war,Vietnam War	1313150742,1313150742
261	1680	3.5	1313072247	fun	1313072247

« 1 »

```
In [10]: review[review.tags.contains('funny')].rating.value_counts()
```



扫码观看大会视频

连接生态能力：IntelliJ IDEA – MaxCompute Studio

IntelliJ 是一个非常流行的 IDE 产品，支持 Java、Python 等多种编程语言，提供丰富的辅助功能，能显著提升程序员开发效率。

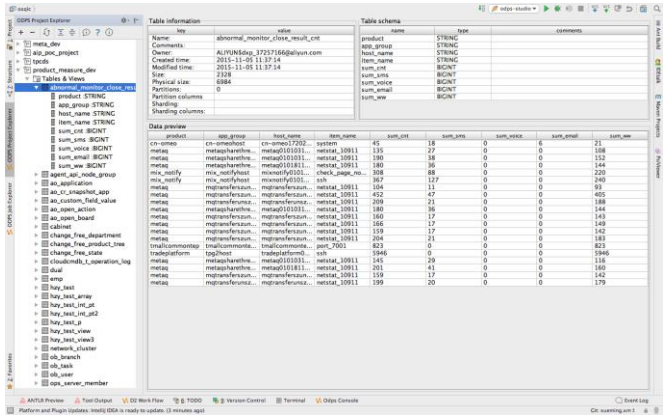
MaxCompute Studio 基于 IntelliJ 平台提供了一套扩展插件，提升 MaxCompute 用户的开发体验。

Eclipse 平台上的插件也在开发中。

MaxCompute Studio

MaxCompute 的本地集成开发环境，提供：

- MaxCompute 项目空间浏览器
- 集成数据上传、下载工具
- 集成 MaxCompute 命令行客户端
- UDF 开发示例代码和工具支持
- SQL 脚本基于语义的语法高亮、智能提示
- 实时语法和类型检查、错误提示
- SQL 作业提交和进度显示
- 历史作业日志查询和可视化



2016 The Computing Conference

THANKS