Challenges and Opportunities of Apache Flink® Ecosystem

公司：阿里巴巴

职位：资深技术专家

演讲者：王绍翾 (大沙)

公司：阿里巴巴

职位：高级技术专家

演讲者：章剑锋 (简锋)

王绍翾（花名“大沙”）
Shaoxuan Wang

阿里巴巴资深技术专家
Senior Staff Engineer at Alibaba

shaoxuan.wsx@alibaba-inc.com
shaoxuan@apache.org

北京大学
EECS

美国加州大学圣地亚哥分校
Computer Engineering

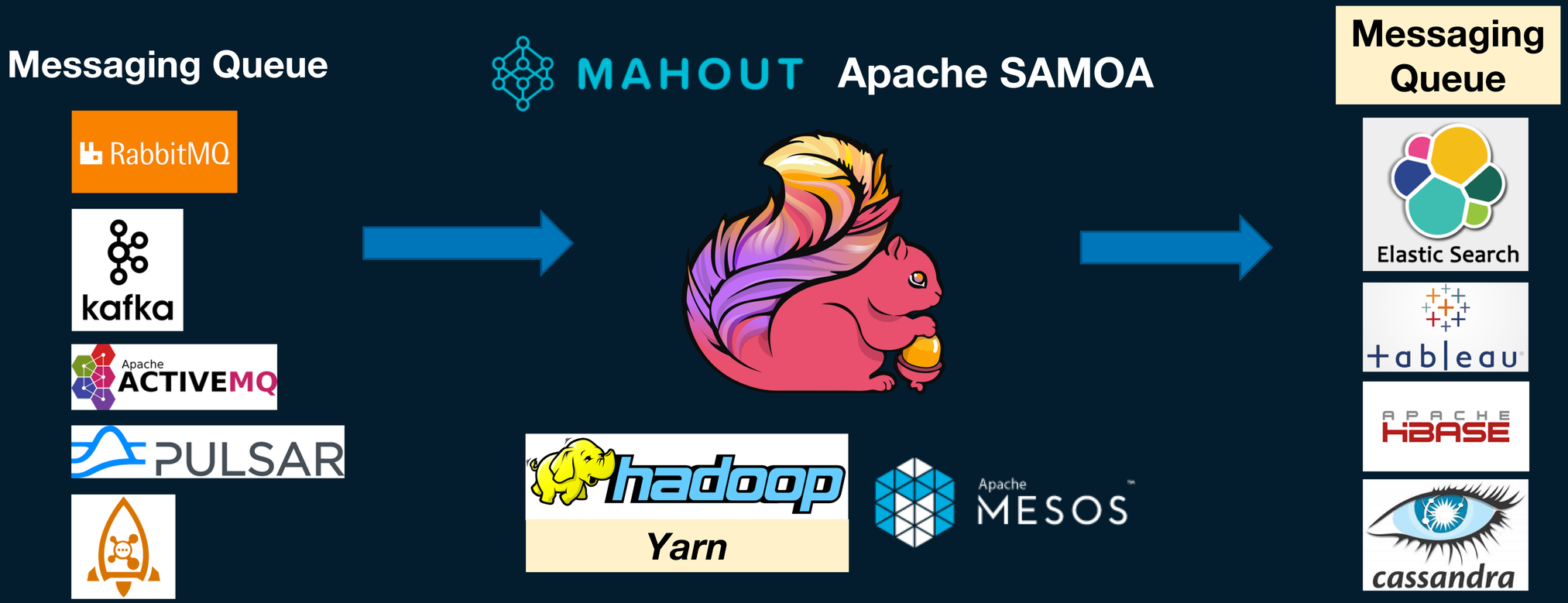
博通（Broadcom）
High-Perf Platform

脸书（Facebook）
Social Graph Storage

阿里巴巴
Real-Time Data Infra

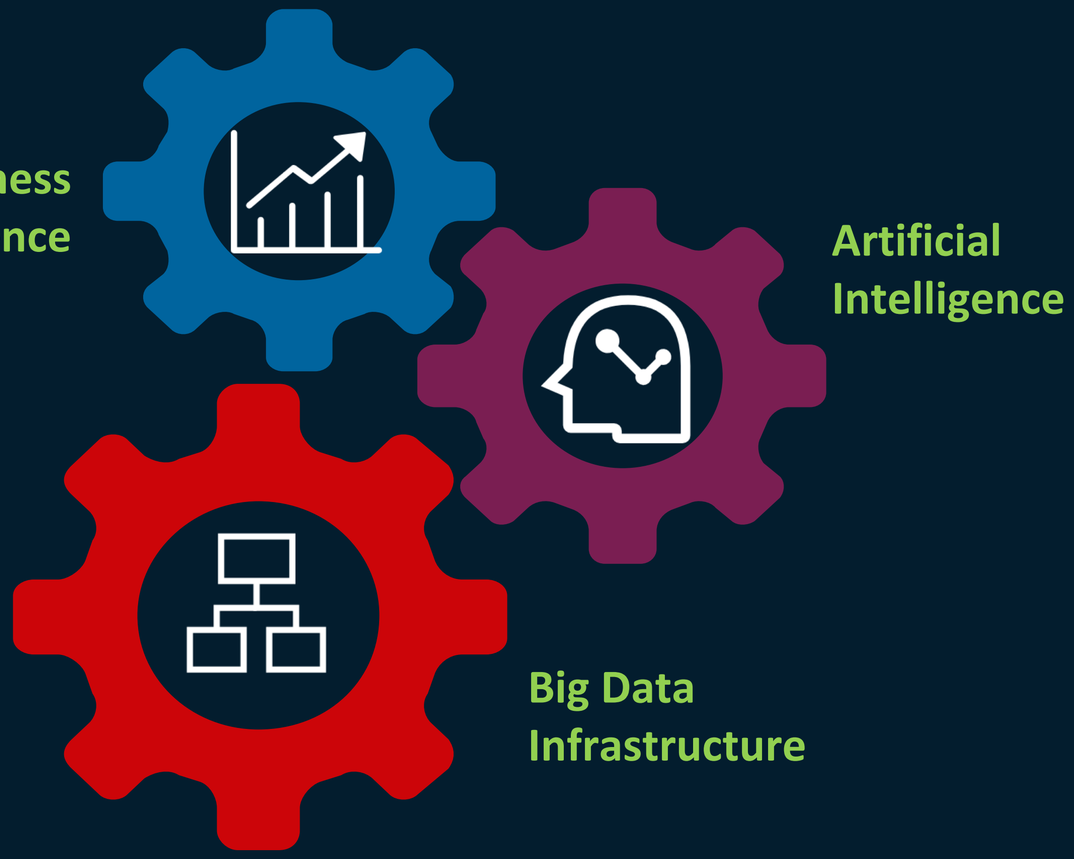
负责大数据实时计算平台，算法工程，和技术生态

Apache Flink Ecosystem - Present



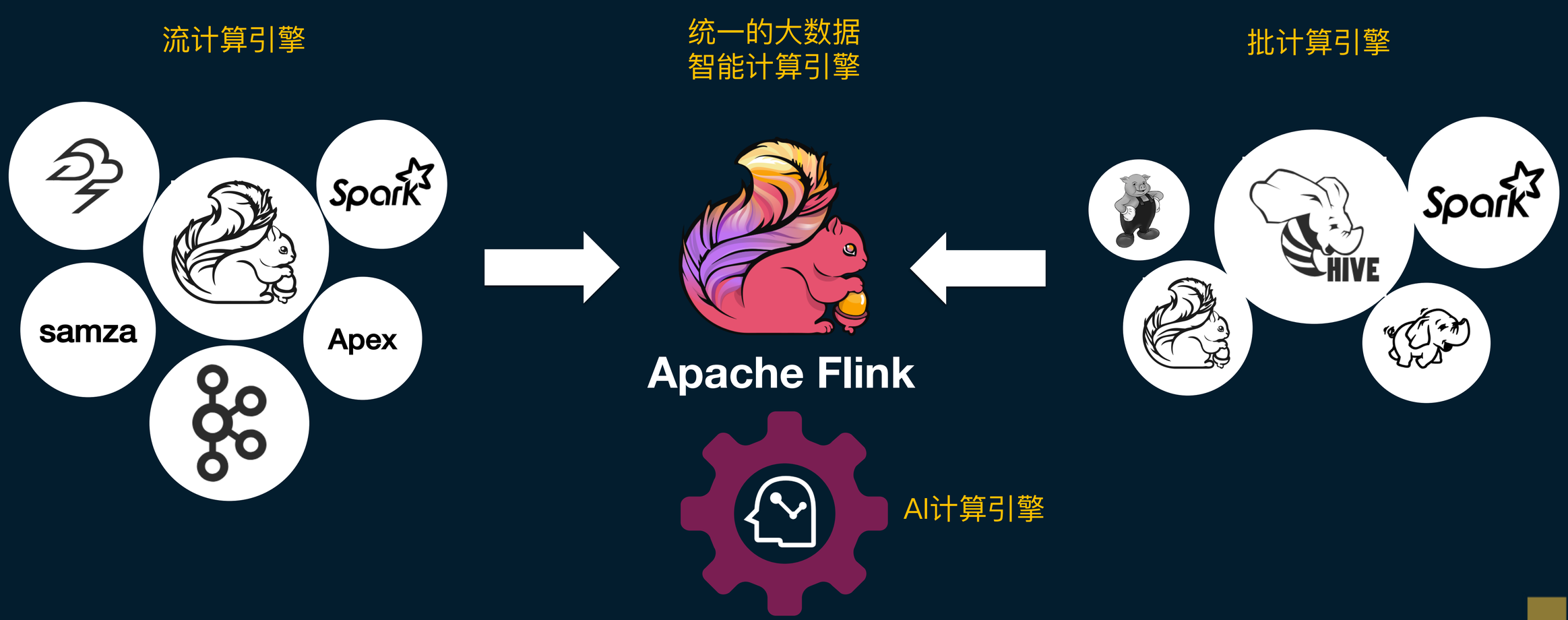
Apache Flink is the most sophisticated open-source Stream Processor

Intelligent Big Data Computing

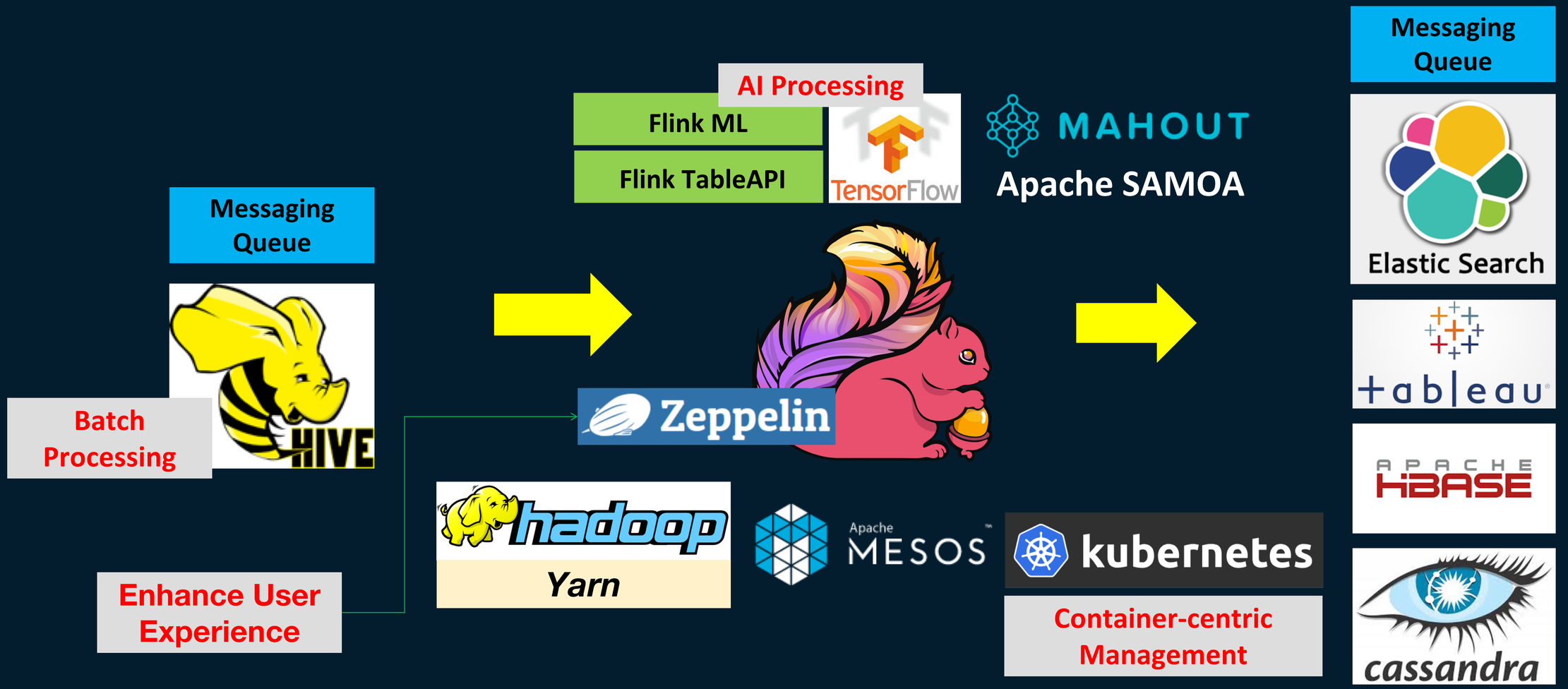


Can Apache Flink become an unified engine for intelligent big data computing?

Build Intelligent Big Data Platform with Apache Flink



Apache Flink Ecosystem – Future



Flink — Powerful and Unified Compute Engine

FLINK FORWARD

API的统一:
unified batch and stream processing, same query, same results

Query Processing的统一:
unified query optimization and query execution framework

应用的统一:
switch between batch processing and streaming seamlessly

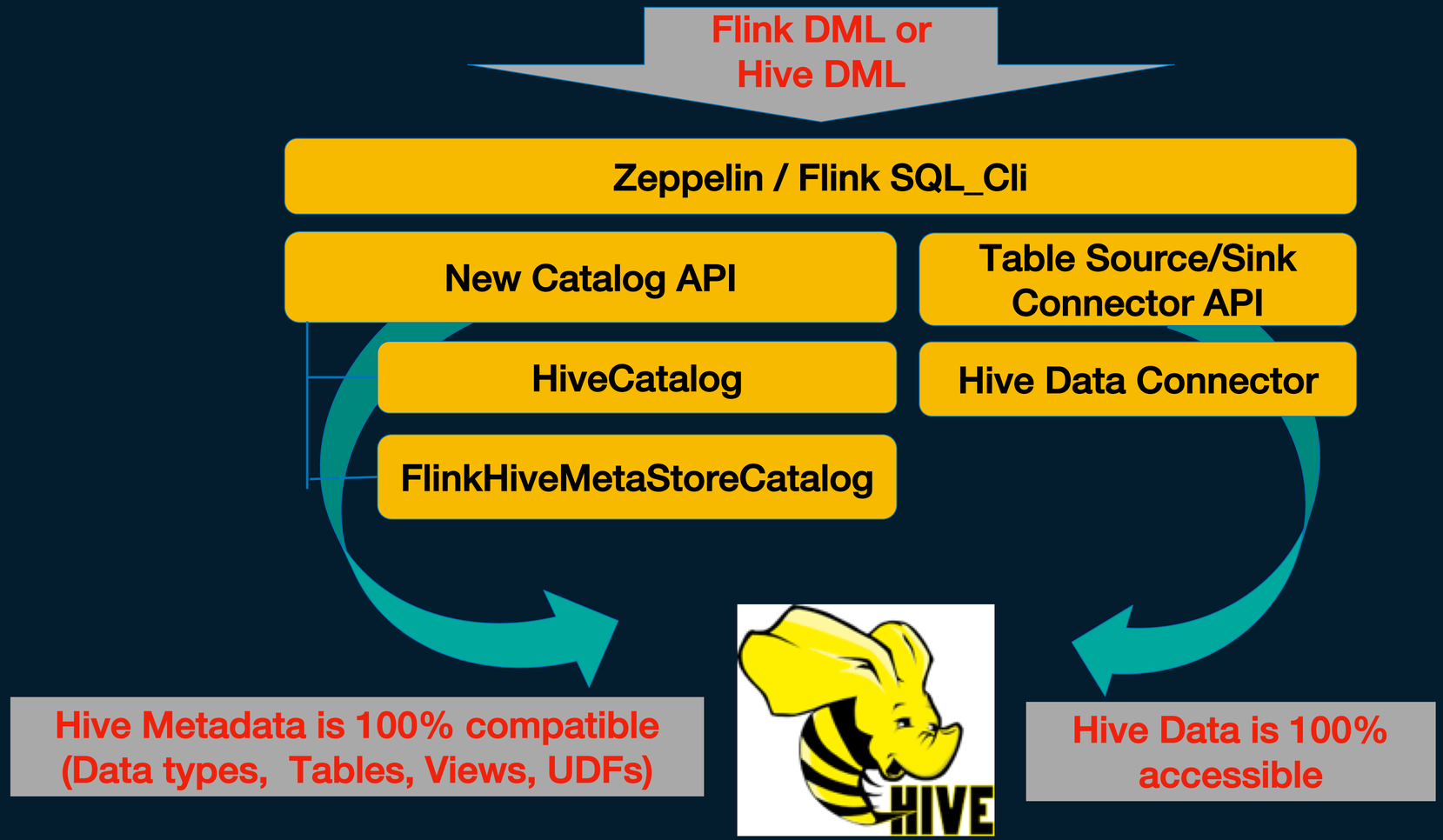
Can obtain 10x performance improvement in batch processing with recent optimization

The Practice and Challenges in Building a Stream-Based Unified Big Data Processing Engine

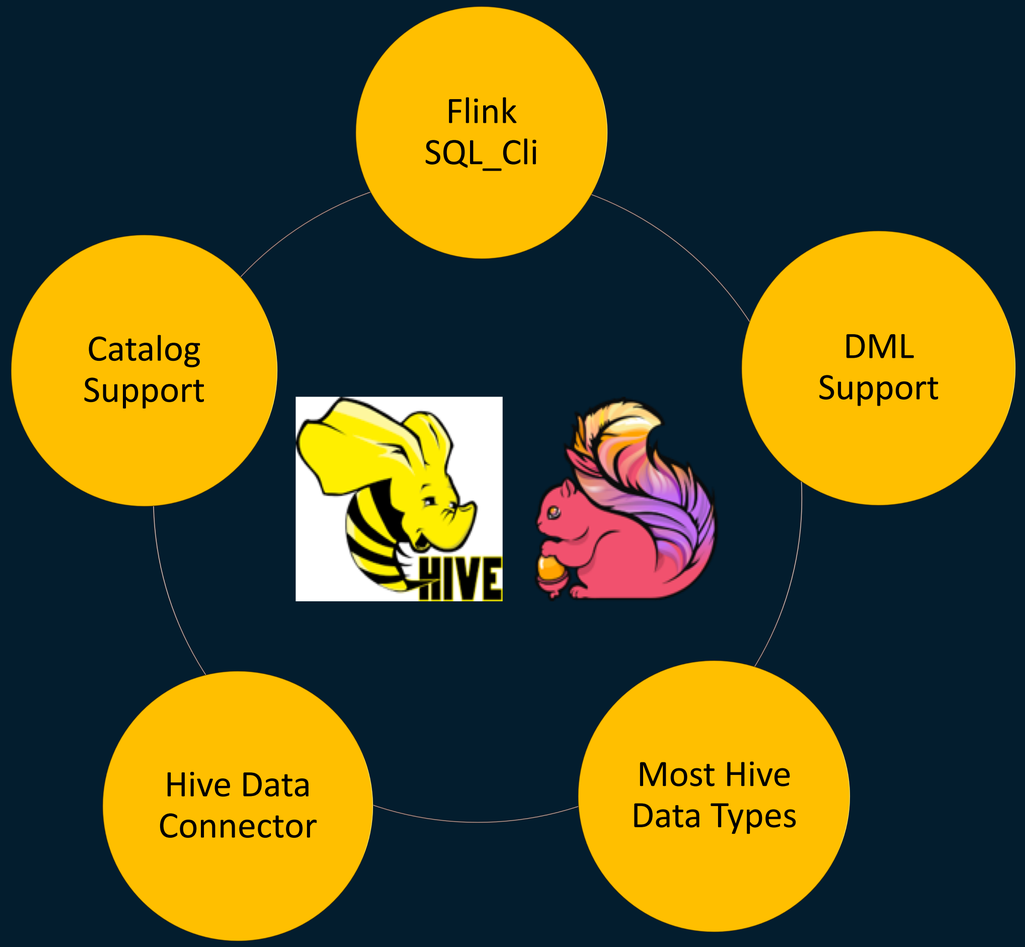
Kurt Yang, Staff Engineer at Alibaba
Jark Wu, Senior Software Engineer at Alibaba

3:20 – 4:00pm
分会场三 310

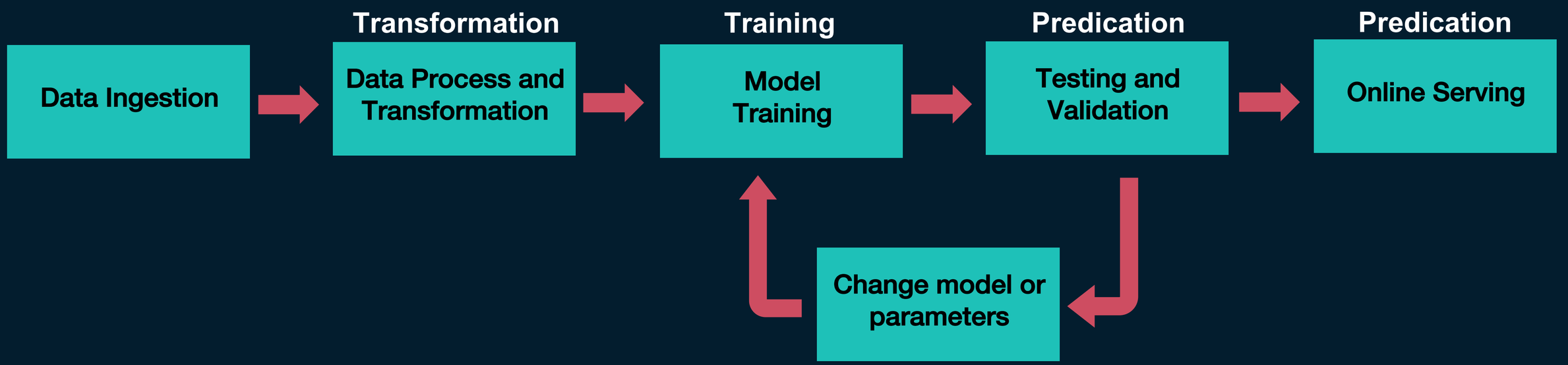
Integrate Flink with Hive Ecosystem (FLINK-10556)



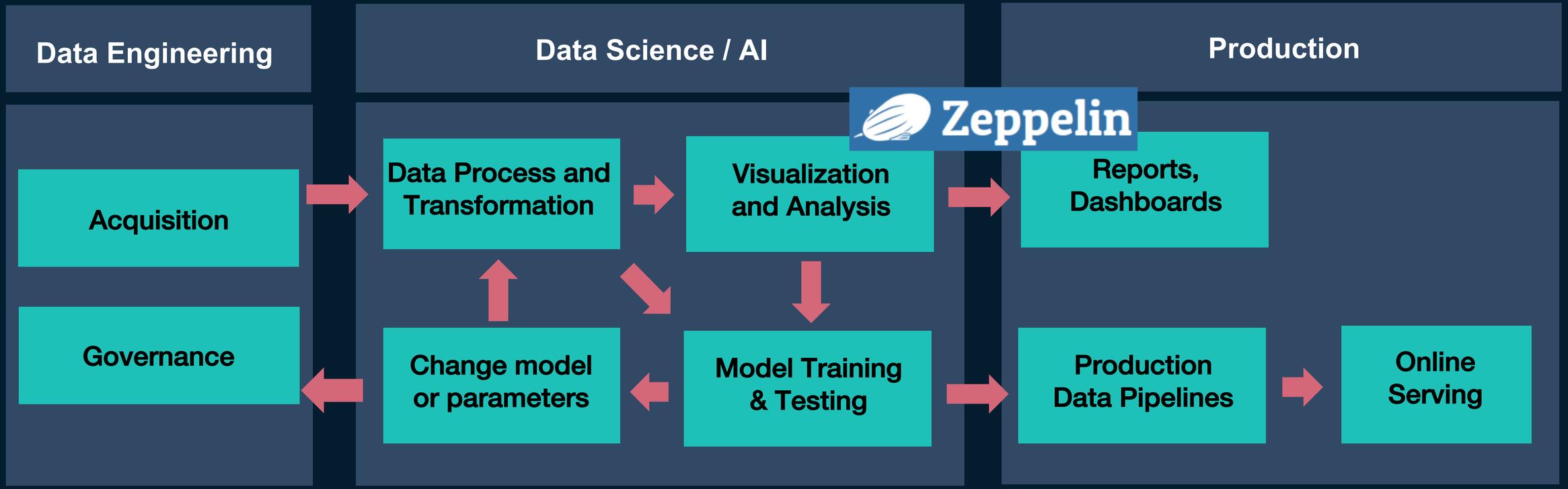
Integrate Flink with Hive Ecosystem (FLINK-10556)



AI Processing



AI Processing



TableAPI for AI Processing

算法实验
Ad-hoc Algorithm Experiments

迭代收敛
Iterate until converge

基于行的计算
Row-based processing

批计算+流计算
Batch + Stream processing

动态模型更新部署
Dynamic model update and deployment

交互式编程 (Flink-11199)
Interactive programming

迭代计算 (coming soon)
Iterative processing

基于整行的计算 (FLIP29, Flink-10972)
Row-based API

批流统一的API
Unified API for batch and stream processing

Simplify Machine Learning With Apache Flink® TableAPI

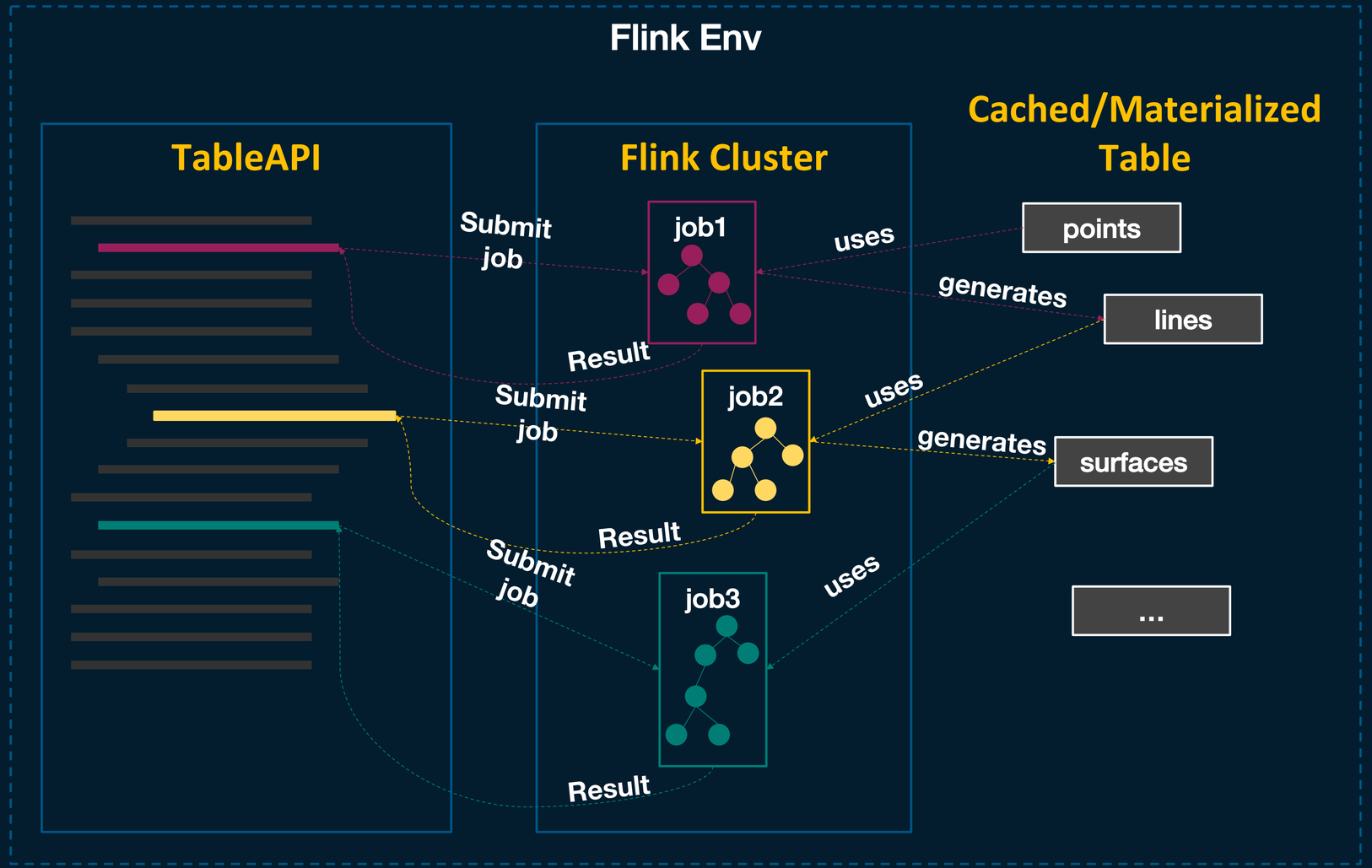
Jiangjie Qin, Staff Engineer at Alibaba

Jincheng Sun, Staff Engineer at Alibaba

2:40 – 3:20pm

分会场一 309B

Interactive Programming (Flink–11199)



Introduce Row-based TableAPI (FLIP29, Flink-10972)

UDF	Single Row Input	Multiple Row Input
Single Row Output	ScalarFunction	AggregateFunction
Multiple Row Output	TableFunction	*TableAggregateFunction (new)*

Table Method	Single Column Output	Multiple Column Output
ScalarFunction	Table.select	*Table.map*
AggregateFunction	Table.select	*GroupedTable.agg*
TableFunction	N/A	*Table.flatmap*
TableAggregateFunction	N/A	*GroupedTable.flatagg*

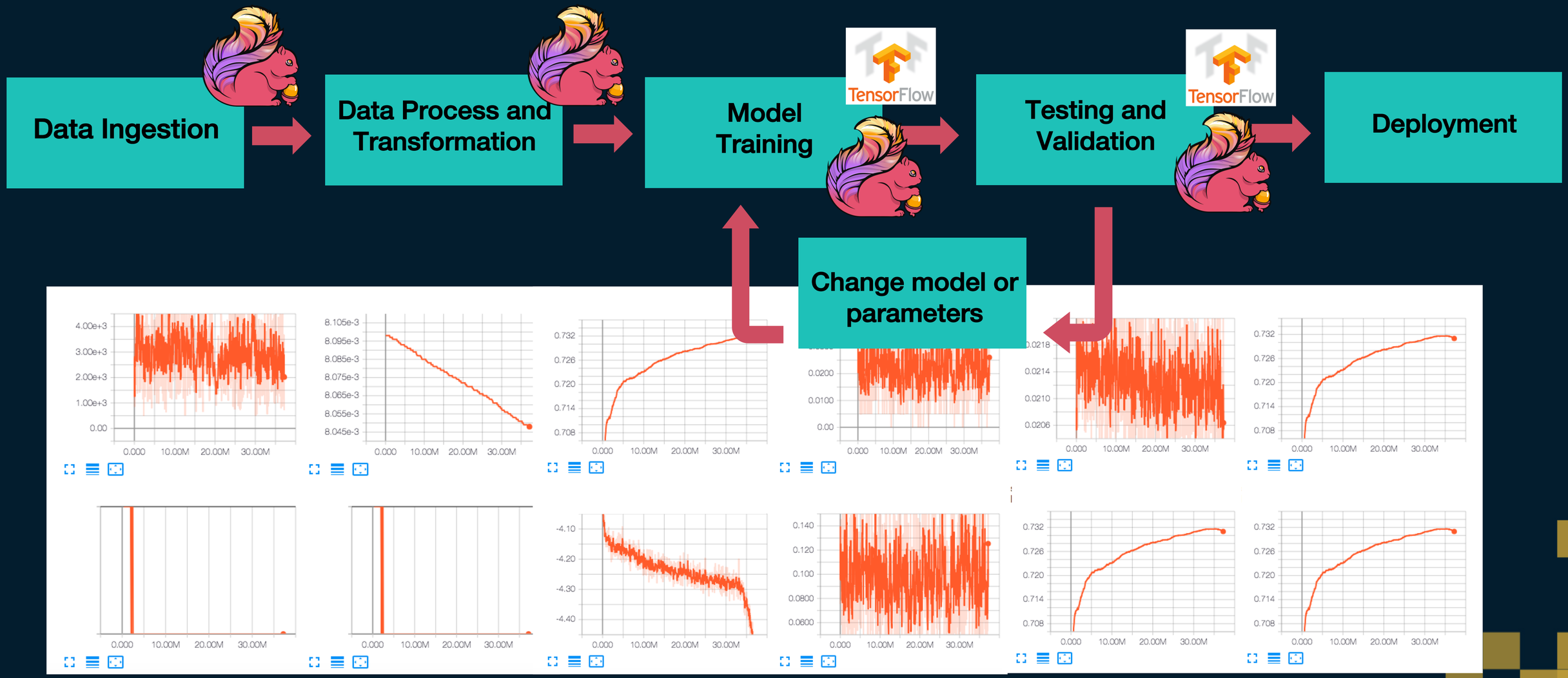
Improve Flink ML Pipeline (Flink-11095)

embracing TableAPI：
更容易理解和上手；批流统一；ETL处理和AI计算可以无缝集成

support streaming ML:
支持实时训练，实时预测

provide a better AI experience:
提供更好的开发工具，以及更好的平台化管理工具

Tensorflow on Flink



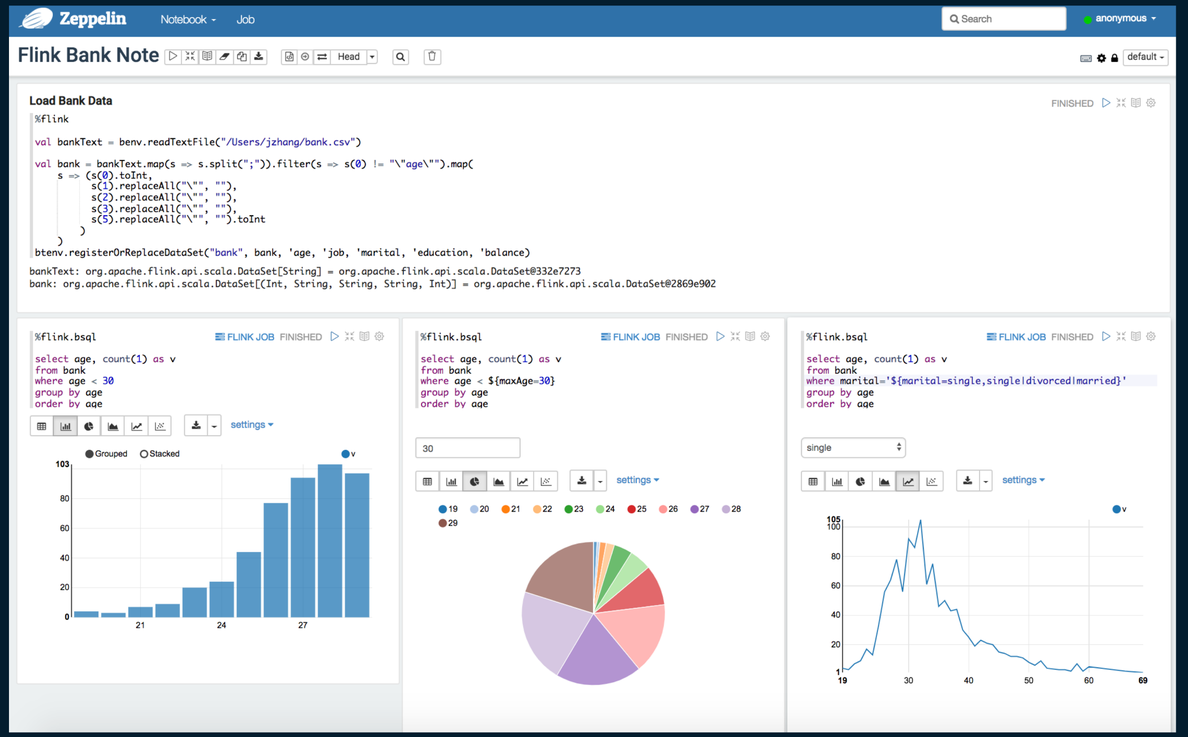
About Me — 章剑锋 (Jeff Zhang)

- Joined Alibaba in 2018/10
- Worked in Hortonworks，Pivotal & eBay
- Apache member， PMC of Apache Tez, Livy & Zeppelin

Apache Zeppelin

基于web的交互式分析Notebook, 支持Sql, Scala等等

Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.



Enhance Flink Experience in Apache Zeppelin (Zeppelin-3913)

效率 (Efficiency)

- 快速开发业务逻辑，无需打包，部署，配置等等
  Quick development, no build, no packaging and easy configuration)

- 快速定位错误，无需跳转到Flink集群服务器
  Identify error fast, no need to switch to JM dashboard for exception)

- 高效管理操作Flink Job（Cancel Job， SavePoint 等等）
  Manipulate Flink job easily, including submission, cancel and resume from savepoint)

集成 (Integration)

- 数据可视化 （静态数据 + 动态数据）
  Data Visualization （Static + Dynamic）
- 其他工具的集成（Hive，Pig， Python，Markdown 等） Integration with other data tools （Hive, Pig, Python, Markdown and etc.）
- 机器学习框架的集成（Tensorflow， Keras 等）
  Integration with Machine learning frameworks (Tensorflow, Keras and etc.)

企业级特性（Enterprise Feature）

- 多租户（Multi-tenancy）
- 安全验证 （Authentication & Authorization）
- 项目代码管理 (Project Management)
- 与云服务的集成（Integration with Cloud）

Take Away

Batch Processing

- Unified Batch and Stream Processing
- Integrate Flink with Hive Ecosystem (FLINK-10556)

AI Processing

- Interactive Programming (FLINK-11199)
- Iterative Processing
- Row-based TableAPI (FLIP29, FLINK-10972)
- Tensorflow on Flink
- Flink ML pipeline (FLINK-11095)

Zeppelin

- Enhance Flink Experience in Apache Zeppelin (ZEPPELIN-3913)

Demo

- Query Hive Meta and Data via Flink SQL_CLI
- Flink Batch and Stream Processing on Zeppelin


THANKS