Table of Contents

Sparkling: The Tencent Cloud Data Warehouse
MemVerge PMEM Centric Elastic Spark Solution
Performance
Ongoing Work
Summary
Sparkling Cloud Data Warehouse

- A PB scale elastic data warehouse
  - fast deployment, resource elasticity, performant, cost-effective
Sparkling: Architecture Overview

Sparkling Data Warehouse Manager

Query Engine
- JDBC
- Livy-Thrift
- SQL Optimizer
- Execution Engine
- Spark
- Columnar Storage
- Data Source Connector

DB IDE and Notebook
- DB IDE
- Query Monitor
- Workflow Mgmt
- DS Notebook

Data Application
- Analysis
- BI
- ML

Tencent Cloud Infrastructure

Data Lake
- RDBMS
- COS
- NoSQL
- ES
- KAFKA
- ...

Data Catalog and Management
- MetaStore
- Data Catalog
- Atlas
- RDBMS
Sparkling: Hiring and Evolving Open Source Technologies

- Flume
- Livy
- Parquet
- HDFS
- Hive Metastore
- YARN
- ORC
- Airflow
- Zeppelin
- ATLAS
- Sqoop

......
SparkSQL

- **Outstanding query engine**
  - Open source with large and active community
  - Fully compatibility with ANSI SQL 2003
  - High performance

- **Integration with Spark ecosystem**
  - Spark Streaming
  - GraphX
  - MLlib/ML

By Michael Armbrust @ Databricks
Three types of nodes: Master, Core and Elastic

Scale in/out for elastic nodes

YARN

NM

NM

RM

NM

Elastic Node

Core Node

Core Node

Master Node

Scale in/out

HDFS

DN

DN

NN

Elastic Node
However, Elastic story is not so easy…

- Ideally, compute node should be stateless

- Actually, it is not…
  - Node decommission will bring down running tasks
  - Map output get lost during node decommission
Key solution: Gracefully Decommission

- Add new state of NM – “decommissioning”
- No new containers get launched in decommissioning NMs
- Decommission nodes if all running containers are finished or timeout
- Umbrella JIRA: YARN-914
Key Solution: Independent Shuffle Service

- Shuffle I/O are **decoupled** from a specific network/storage.
- Shuffle read and write can be implemented using **configurable** network transports and backend storage.
- Listeners are inserted into different stages of the shuffle to apply hooks.
Sparkling: Evolving from Data Warehouse to Data Lake

- BI
- Machine Learning
- Data Management
- Data IDE

Sparkling Data Lake

- Database
- Data Warehouse
- Object Store
- NoSQL
MemVerge Elastic Spark Solution

- Spark Cluster
- Spark Accelerator
- Ethernet Switch
- RDD Caching and Storage
- Data Source
- Shuffle Data
Intel® Optane™ DC Persistent Memory
Spark Integration

RDD Caching and Storage

Spark with additional RDD persist APIs

Data Source

Hadoop compatible storage APIs

Shuffle Data

A new generic shuffle manager

MemVerge Data Platform
MemVerge Splash Shuffle Manager

- A flexible shuffle manager
  - supports user-defined storage backend and network transport for shuffle data

- Open source
  - https://github.com/MemVerge/splash

- Spark JIRA: SPARK-25299
PMEM Centric Data Platform

MemVerge Spark Adaptors

MemVerge SDK

Cluster Shared Persistent Memory

Compute Node 1
- DRAM
- PMEM

Compute Node 2
- DRAM
- PMEM

Compute Node 3
- DRAM
- PMEM

Compute Node 4
- DRAM
- PMEM

... (continues)

Compute Node N
- DRAM
- PMEM

© 2019 Storage Networking Industry Association. All Rights Reserved.
Source Data in Remote PMEM

HiBench Wordcount Time

10min

SSD

2.2min

PMEM

*5 Spark compute node, 1 remote PMEM node, Data size: 610GB
Persisting RDD to Remote PMEM

*10 compute nodes, 1 remote PMEM node, production analytics workload
Shuffling with Remote PMEM

*10 compute nodes, 1 remote PMEM node, production data warehouse workload
Ongoing Work

- TPC-DS performance study

- Better cloud readiness
  - Virtual machine support
  - Container support
Summary

- PMEM will bring fundamental changes to ALL data centers and enable a data-driven future
- MemVerge and Tencent Cloud deliver better scalability and performance at a lower cost not just for Spark
  - AI, Big Data, Banking, Animation Studios, Gaming Industry, IoT, etc.
  - Machine learning, analytics, and online systems
- Thank you Intel for supporting our work!

Junping Du: junpingdu@tencent.com
Yue Li: yue.li@memverge.com