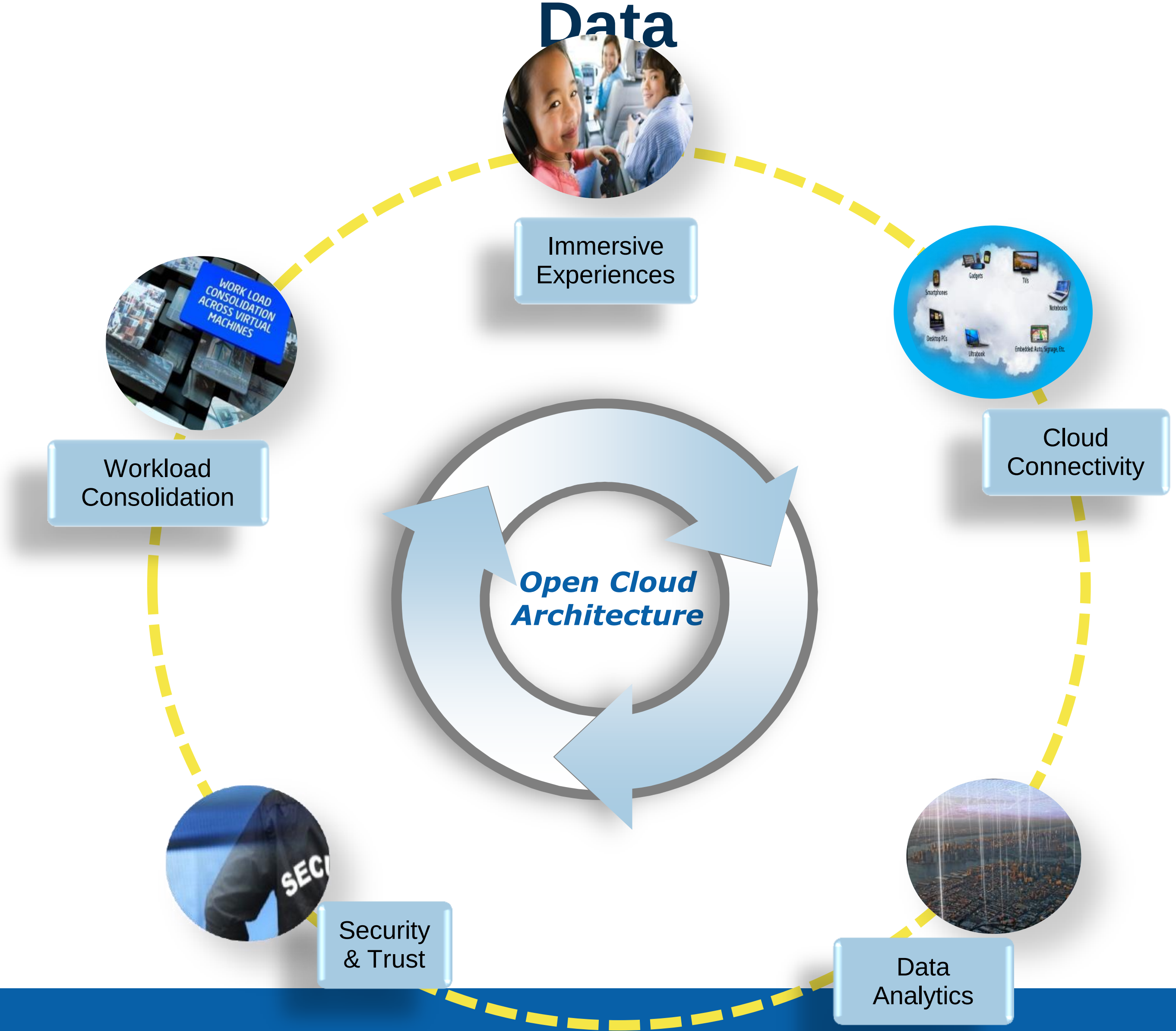
Hadoop: the Intel Way (Hadoop的英特尔之道)

Bring New Analytics Capabilities to Hadoop Stack

何京翔
英特尔亚太研发有限公司总经理

Software and Services Group

Cloud and IOT: More Users, More Device, More Data


WORK LOAD CONSOLIDATION ACROSS VIRTUAL MACHINES
Gadgets
TVs
Smartphones
Notebooks
Desktop PCs
Ultrabook
Embedded: Auto, Signage, Etc.
SECU

Immersive Experiences

Cloud Connectivity

Data Analytics

Security & Trust

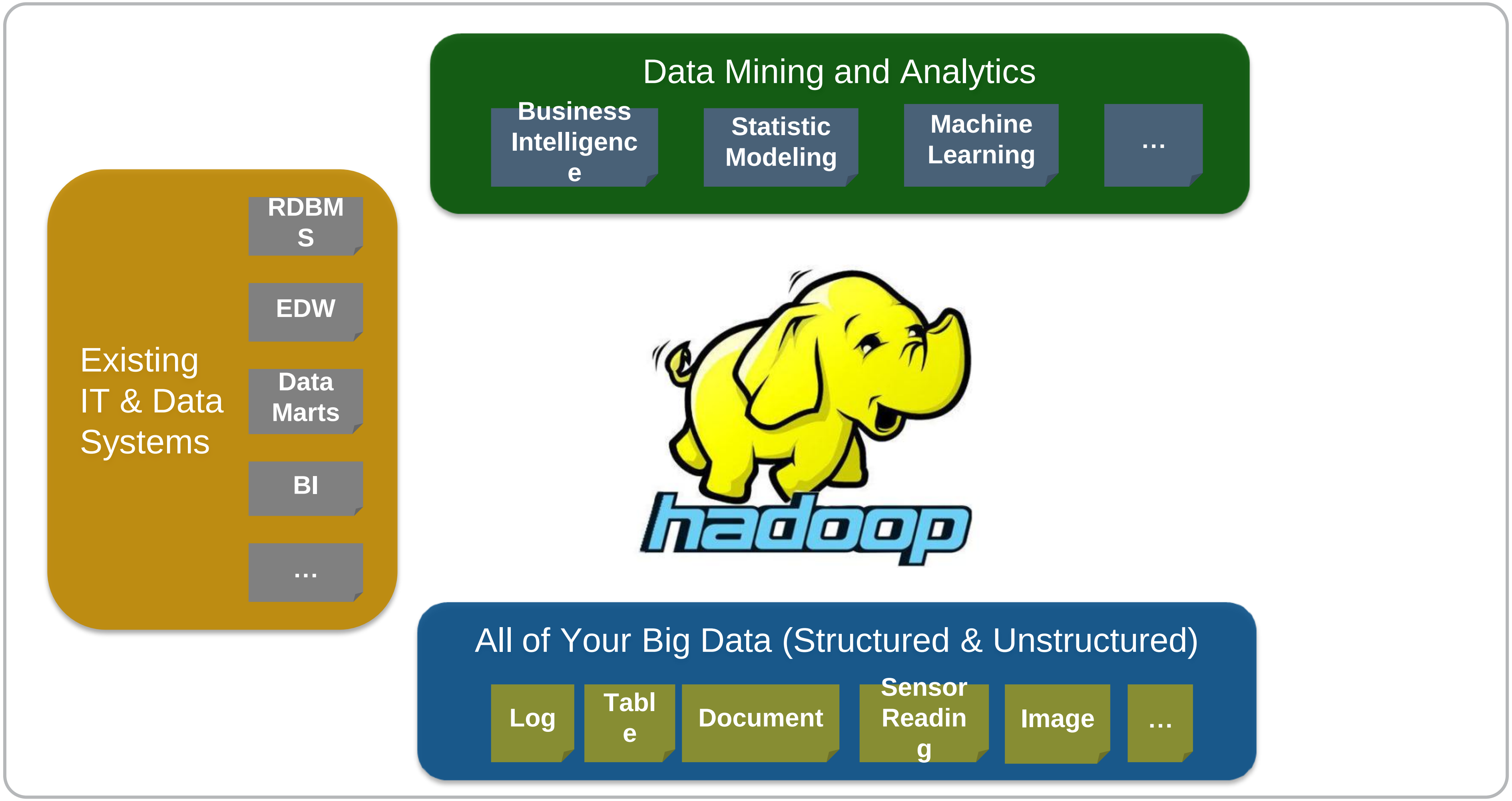
Workload Consolidation

Open Cloud Architecture

Intel's Vision

This decade we will create and extend computing technology to connect and enrich the lives of every person on earth

Our Big Data Goal: Make *Hadoop* the Foundation of Next-Gen Data Analytics Platform



Data Mining and Analytics

Business Intelligence

Statistic Modeling

Machine Learning

…

Existing IT & Data Systems

RDBMS

EDW

Data Marts

BI

…

All of Your Big Data (Structured & Unstructured)

Log

Table

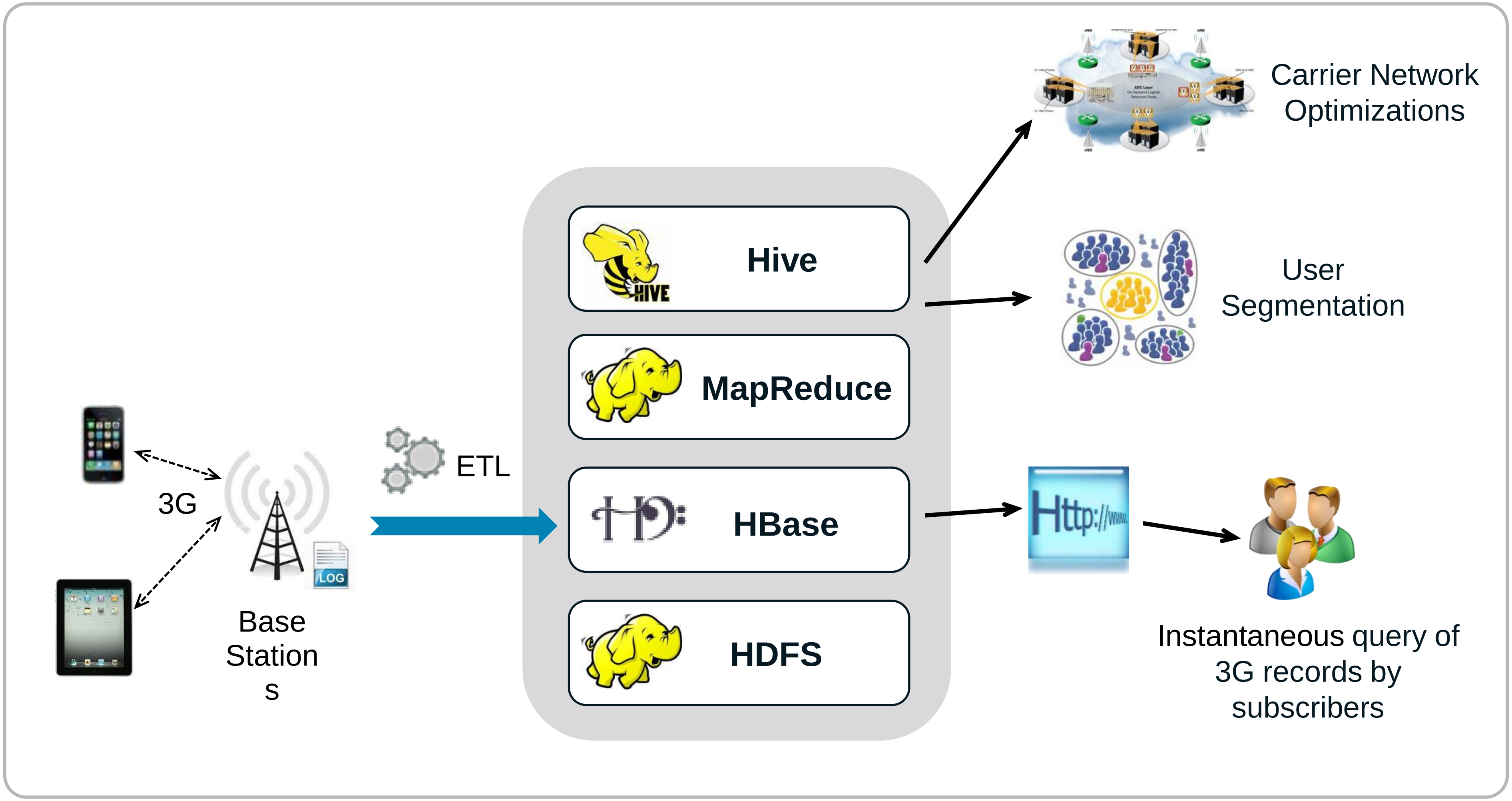
Document

Sensor Reading

Image

…

Hadoop in Telecom



3G

Base Station s

ETL

Hive

MapReduce

HBase

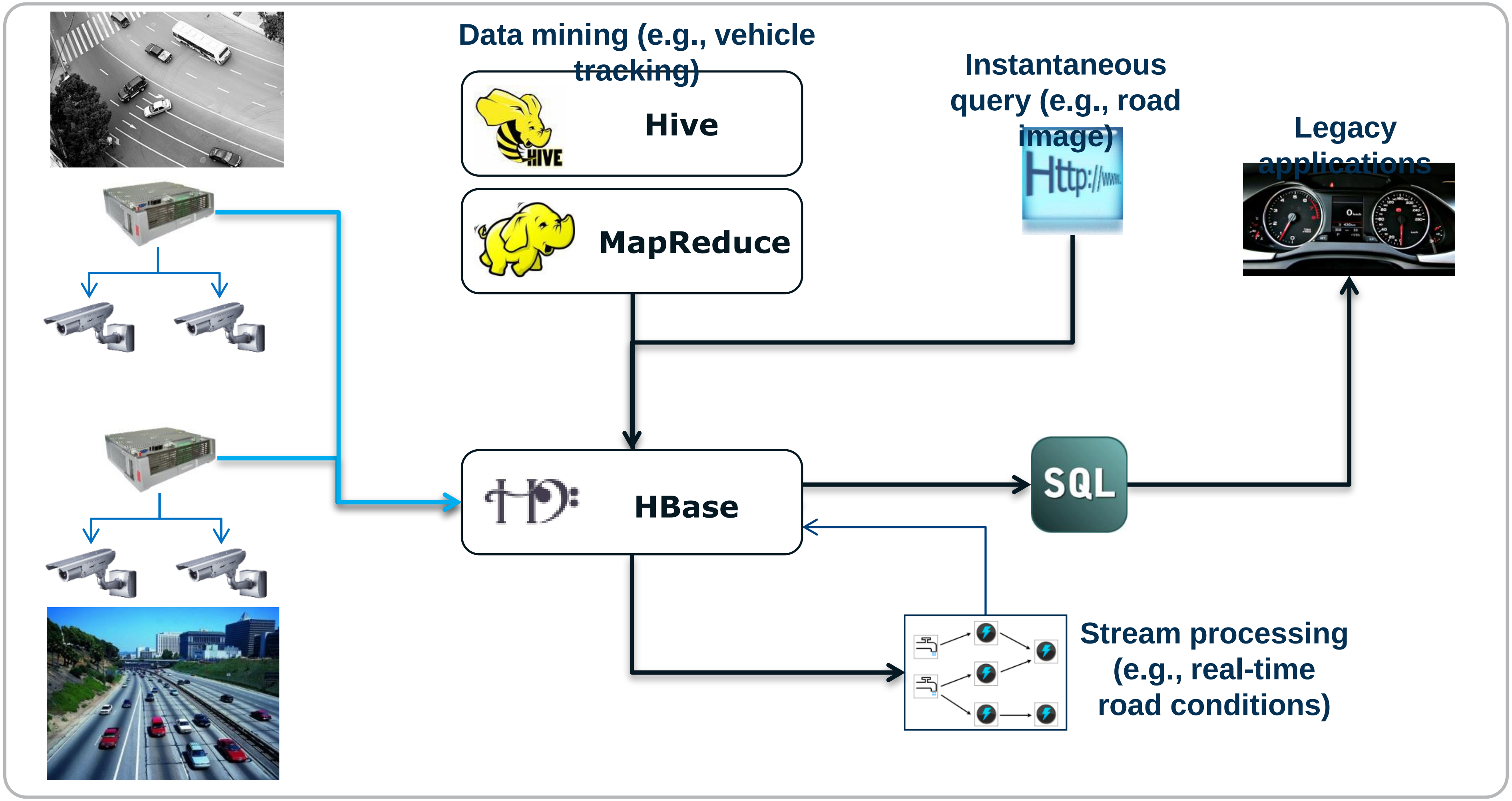
HDFS

Carrier Network Optimizations

User Segmentation

Instantaneous query of 3G records by subscribers

Hadoop in Smart City



Data mining (e.g., vehicle tracking)

Hive

MapReduce

Instantaneous query (e.g., road image)

Legacy applications

HBase

Stream processing (e.g., real-time road conditions)

Hadoop的英特尔之道

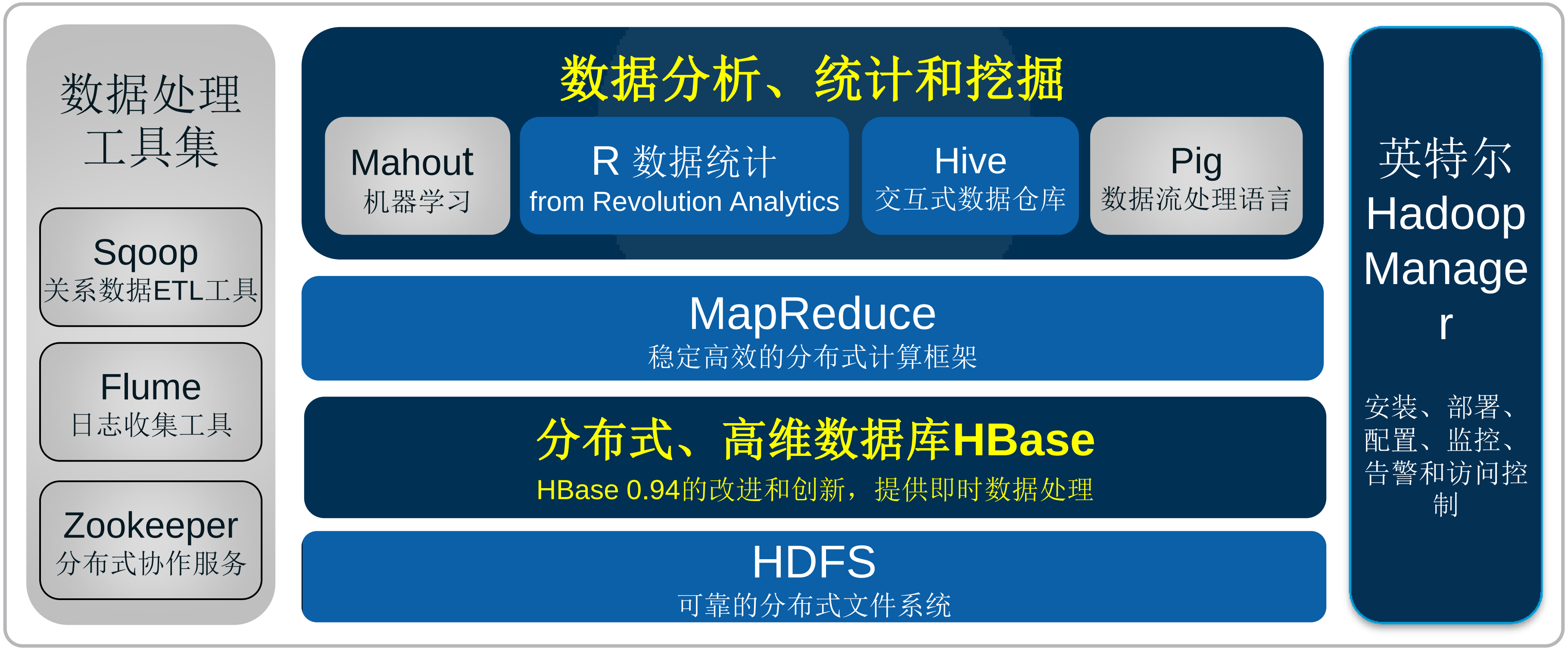
	企业级解决方案 Enterprise-Grade Solution	前沿技术开发 Advanced Development
即时分析 (Instantaneous Analysis) 更易用 (Reduced Complexity) 更高效 (Improved Efficiency)	***英特尔Hadoop发行版*** • 稳定的企业级软件产品 • 针对垂直行业的功能增强	***“Project Panthera”*** • Advanced development and path-finding • Open source and community driven

Bring New Analytics Capabilities to Hadoop Stack

英特尔Hadoop发行版

优化的大数据处理软件产品

稳定的企业级Hadoop发行版	利用硬件新技术进行优化
为Hadoop提供即时数据处理能力	针对行业的功能增强，应对不同行业的大数据挑战



数据处理工具集

Sqoop
关系数据ETL工具

Flume
日志收集工具

Zookeeper
分布式协作服务

数据分析、统计和挖掘

Mahout
机器学习

R 数据统计
from Revolution Analytics

Hive
交互式数据仓库

Pig
数据流处理语言

MapReduce
稳定高效的分布式计算框架

分布式、高维数据库HBase
HBase 0.94的改进和创新，提供即时数据处理

HDFS
可靠的分布式文件系统

英特尔Hadoop Manager

安装、部署、配置、监控、告警和访问控制

“Project Panthera”

Open source initiatives to enable advanced analytics capabilities on Hadoop

https://github.com/intel-hadoop/project-panthera

SQL engine for Hive/MapReduce

- Better integration with existing infrastructure using SQL

Document store on HBase

- Document semantics & significantly speedup query processing on HBase

- Efficient utilization of new HW platform technologies

…

即时分析 (Instantaneous Analysis)

Instantaneous analysis with greatly enhanced HBase

- Stream new data into HBase for analysis in real time
- Support high update rate workloads (to keep the system always up to date)
- Allow very low latency, online data serving
- Etc.

Interactive Query on HBase (英特尔Hadoop发行版)

10X faster than MapReduce

For certain queries on HBase (e.g., group-by aggregation)

HBase Query Engine Layer

- Fast, distributed aggregations directly inside HBase
- Parallel scanning over multiple regions
- Advanced, distributed filtering (CRC32 comparator, fuzzy row filter, etc.)

HBase Query Engine as New Hive Backend

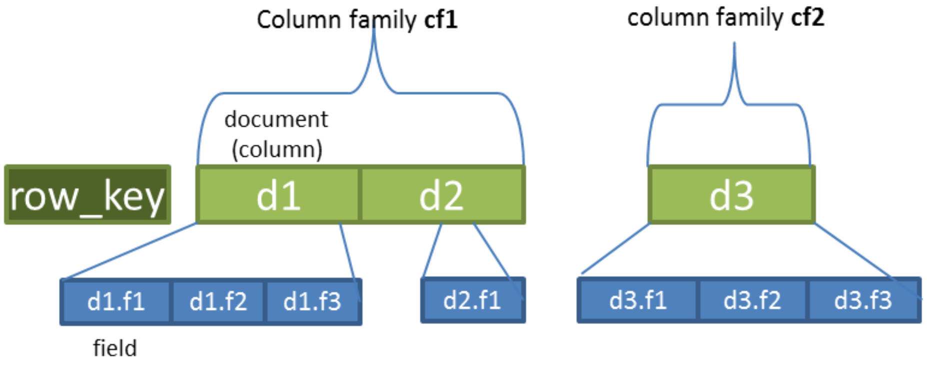
- Most “SELECT” automatically optimized to use HBase Query Engine
  - ✓ “WHERE” using advanced scanner/filter
  - ✓ “GROUP-BY” using distributed aggregations
- “JOIN” stills go to MapReduce

A Document Store on HBase (*“Project Panthera”*)

Up-to 3x storage reduction and 3x query speedup

For Hive/MapReduce query processing on HBase

(See https://github.com/intel-hadoop/hbase-0.94-panthera and HBASE-6800)



DOT (Document Oriented Table) on HBase

- Each row contains a collection of documents
- Each document contains a collection of fields
- A document is mapped to a HBase column and serialized using Avro
- Complete transparent to existing HBase applications

更易用 (Reduced Complexity)

- Better data mining and statistics capabilities
  - ✓ Full-text indexing and search
  - ✓ Statistic modeling with R language
- Better integration with existing infrastructures
  - ✓ Geo-distributed datacenters
  - ✓ Full SQL support for OLAP

Full-Text Indexing and Search (*英特尔Hadoop发行版*)

Full-text indexing and near real-time search for advanced data mining

(E.g., log and click stream analysis, healthcare record analysis, etc.)

Incremental full-text indexing on HBase

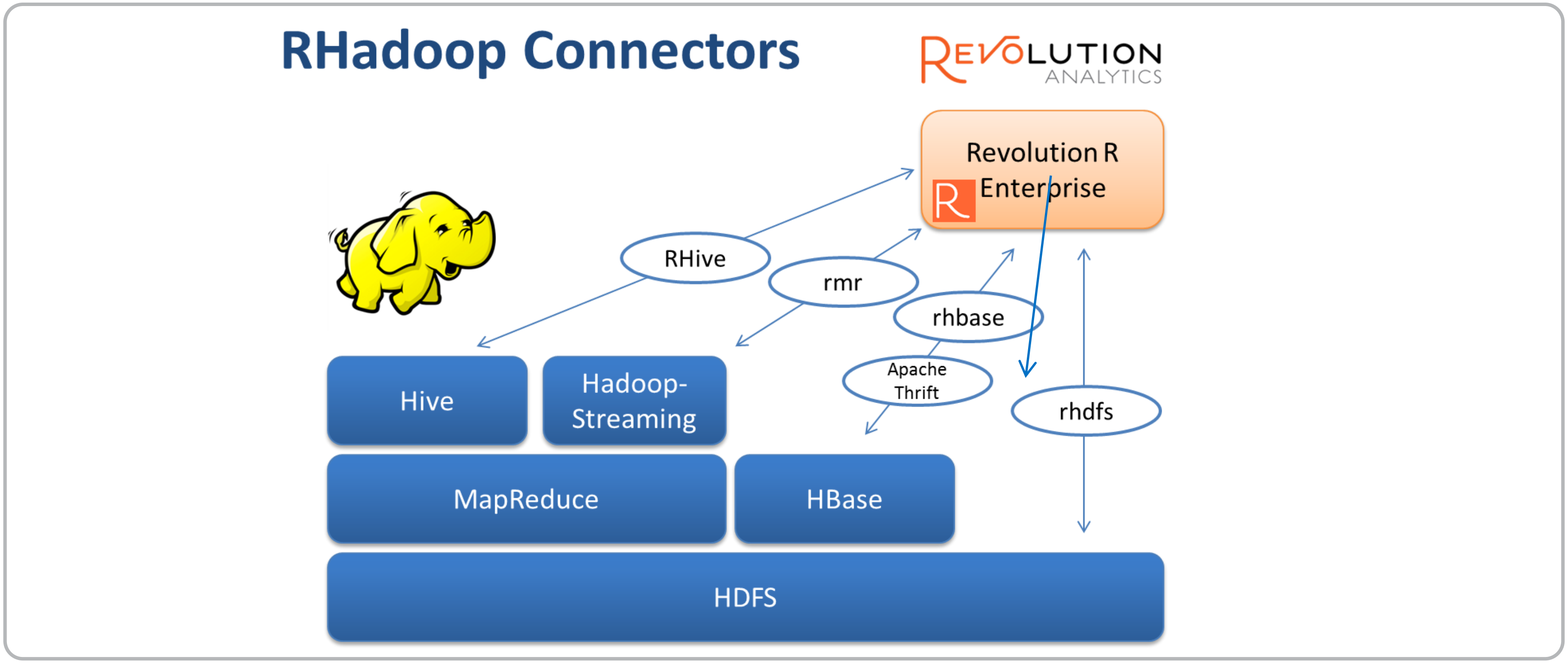
- Full-text indexing for semi-structured data (text, strings, numbers, etc.)
- Index incrementally built when records inserted or updated
- Support very high data insertion / update rate

Near real-time search

- Distributed, keyword or logical expression based search
- Zero delay of searching latest data that are just inserted

Bring R Statistics into Hadoop (*英特尔Hadoop发行版*)

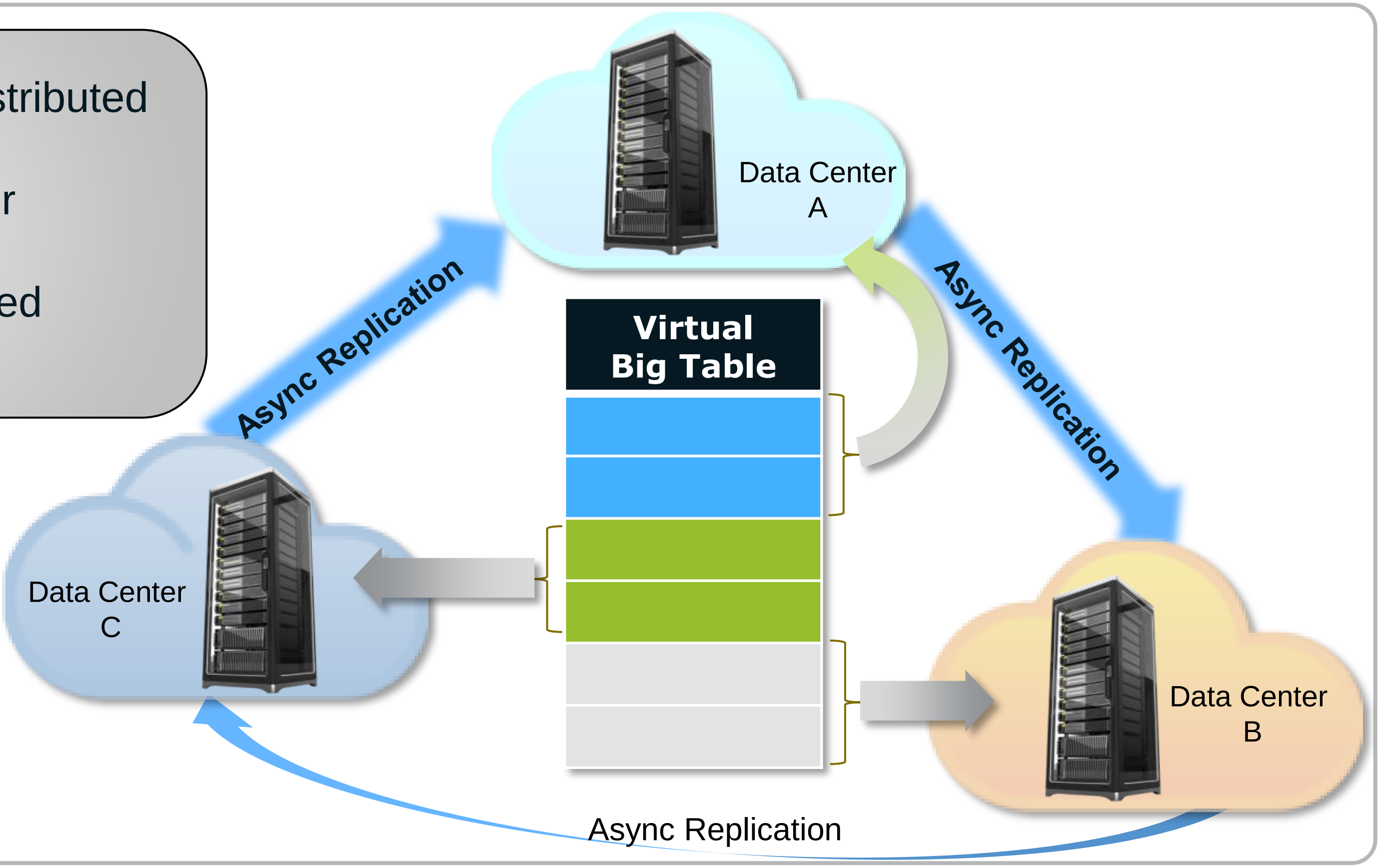
Distributed Statistic Modeling on Hadoop using R language



Cross-Datacenter BigTable/HBase (英特尔Hadoop发行版)

A virtual Big Table overlaid over existing geo-distributed data centers

- Global table view
- Data stored in geo-distributed data centers
- Better locality & higher availability
- Data transfer eliminated through distributed aggregation

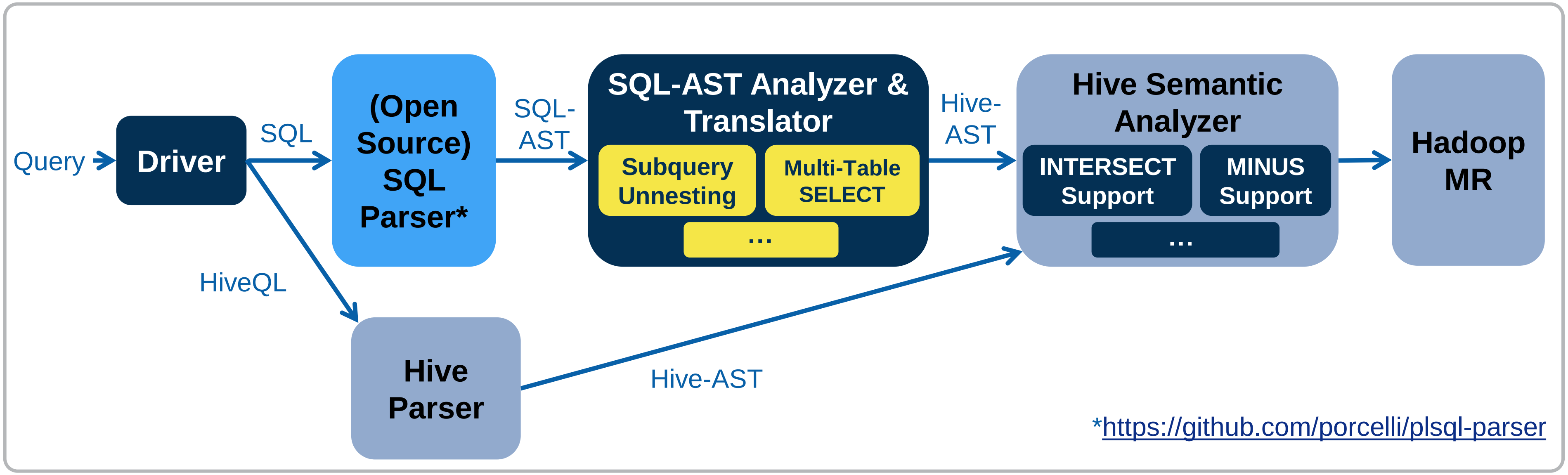


An analytical SQL engine for Hive/MapReduce (*“Project Panthera”*)

Goal: Provide Full SQL support for OLAP in Hadoop

Required by business users, enterprise applications, 3rd party tools (e.g., BI applications), etc.

(See https://github.com/intel-hadoop/hive-0.9-panthera and HIVE-3472)

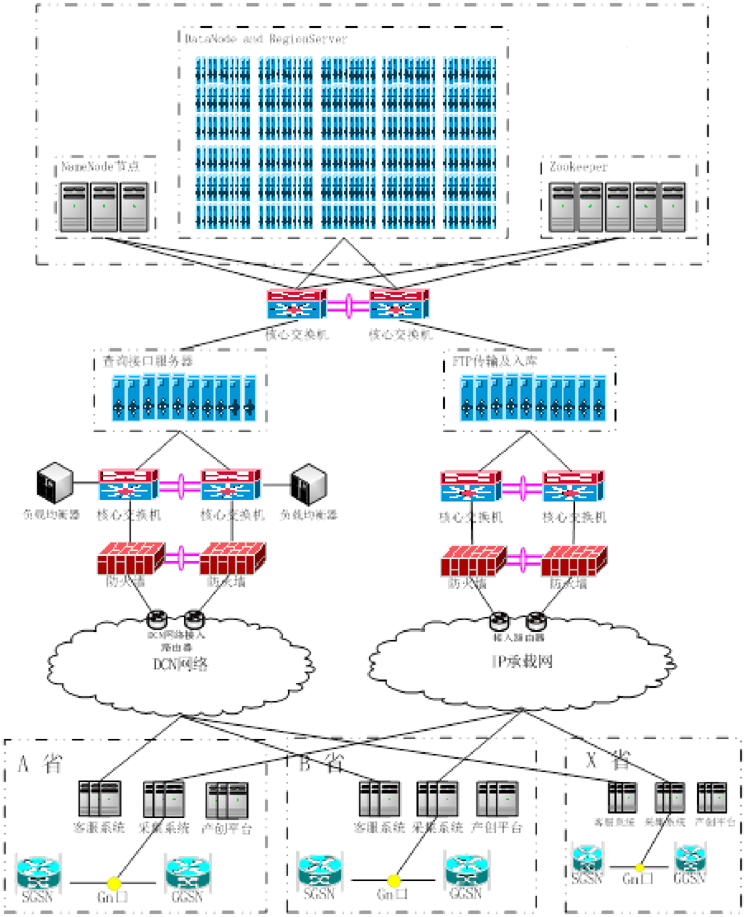


*https://github.com/porcelli/plsql-parser

更高效 (Improved Efficiency)

- Performance benchmarks & tools
- Efficient utilizing of new HW platform technologies (e.g., SSD, infiniband)

*英特尔Hadoop发行版*高效支撑海量移动上网记录分析



联通全国移动用户上网记录查询分析系统

- 国内首个基于Hadoop/HBase的商用电信服务系统
- 系统部署
  - ✓ *英特尔Hadoop发行版*
    - ❖ 满足高性能的数据导入和快速查询。
    - ❖ 稳定、易于部署和管理的企业级方案。
  - ✓ *180+节点*Hadoop/HBase集群
- 系统性能指标
  - ✓ 上网记录入库时间：一般小于*30分钟*，实际约*10分钟*
  - ✓ 具备存储全国移动用户不小于*6个月*的原始上网记录能力
  - ✓ 统计分析的中间报表数据保存不小于*5年*
  - ✓ 上网记录查询速度：不高于*1秒*
  - ✓ 支持并发查询数目：*1000请求/秒*

HiBench & HiTune Performance Tools (*“Project Panthera”*)

HiBench: Hadoop Benchmark Suite

(See https://github.com/intel-hadoop/hibench)

HiBench

1 *Micro Benchmarks*

- Sort
- WordCount
- TeraSort
- Enhanced DFSIO

2 *Web Search*

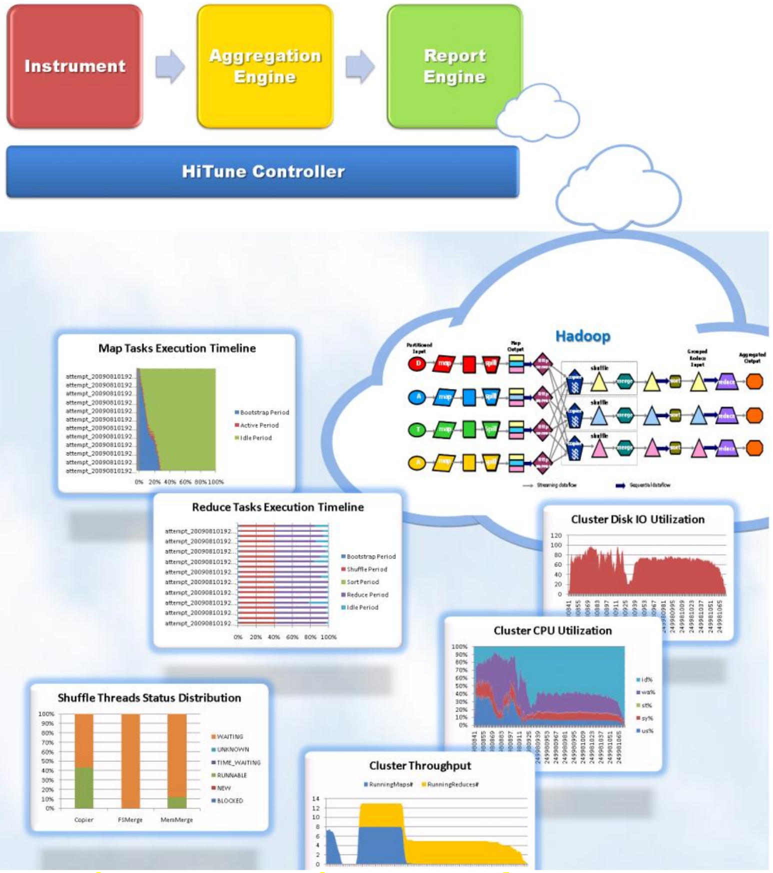
- Nutch Indexing
- Page Rank

3 *Machine Learning*

- Bayesian Classification
- K-Means Clustering

4 *Analytical Query*

- (Hive) Join
- (Hive) Aggregation


Instrument
Aggregation Engine
Report Engine
HiTune Controller
Hadoop
Map Tasks Execution Timeline
Reduce Tasks Execution Timeline
Cluster Disk IO Utilization
Cluster CPU Utilization
Shuffle Threads Status Distribution
Cluster Throughput

HiTune: Hadoop Performance Analyzer

(See https://github.com/intel-hadoop/hitune)

Trying is Believing

英特尔Hadoop发行版免费版 v2.2, 为最终用户和应用提供商提供了一个功能强大、方便易用的大数据入门平台。

- 免费版和企业版共用相同的核心代码
- 免费版包含所有核心增强功能
- 免费版在节点数和系统存储容量上有所限制

英特尔Hadoop发行版主页：www.intel.cn/idh

CSDN英特尔Hadoop发行版社区： http://bbs.csdn.net/forums/intelhadoop

Summary

Immersive Computing = Big Data = Big Opportunities

Intel is committed to deliver better and faster Hadoop solutions for big data analytics

Intel Hadoop Distribution (IHD) Free Edition is here, try it out!

intel
Amazing things happen with Intel inside®