Ingesting Data from Oracle Database 12*c* R2 into a Hadoop System

Building a Data Lake on Linux for Power

Mithun Hr – IBM Systems, India
Wayne T. Martin – IBM Systems, USA
Arun Sar – IBM Systems, USA

Oracle OpenWorld 2017

IBM

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM and Oracle, a shared commitment to Client Value

Sustained Collaboration for Clients
Oracle 30+ years, PeopleSoft 20+ Years, JD Edwards 35+ years, Siebel 15+ years

Mutual Executive Commitment
Dedicated, Executive-led Alliance teams
Senior Executive reviews and functional Management interlocks

More than 100K joint Technology Clients
Hardware and Software support via in-depth certification process

Award-winning Oracle Services Practice
More than 5,500 successful joint services projects, 13,000 skilled Oracle resources worldwide

Vibrant technology Collaboration
Continued joint development delivering Oracle software optimized for IBM hardware, significant skills/resources investment, dedicated International Competency Centers

Cooperative Client Support Process
Dedicated resources and significant program investments

This certifies that

IBM CORPORATION

has achieved the level

ORACLE® Diamond Partner

IBM

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM

POWER8: Designed for data to deliver breakthrough performance

4X
Threads per core

POWER8
SMT8

x86
Hyperthread

Parallel Processing

4X
Mem. Bandwidth[1]

Data flow

POWER8
pipe

x86 pipe

6X
More cache[2] **@
Lower Latency**

x86

POWER8

OpenPOWER™

Moore's Law

Price/Performance

Processor
Technology

x86

Firmware / OS
Accelerators
Software
Storage
Network

POWER8 +
OpenPOWER

2000   2020
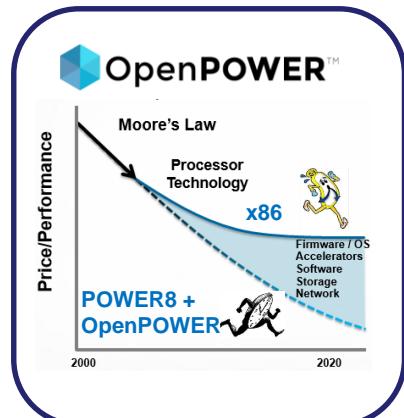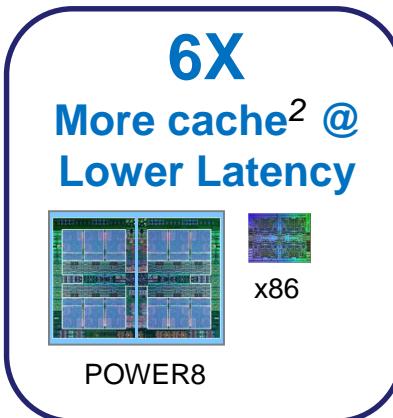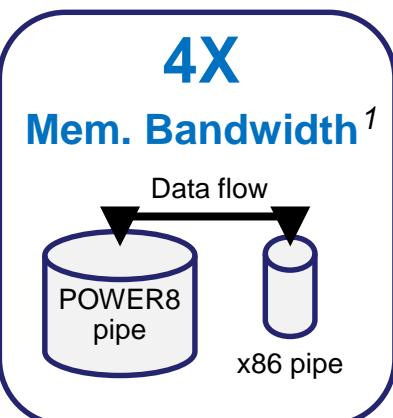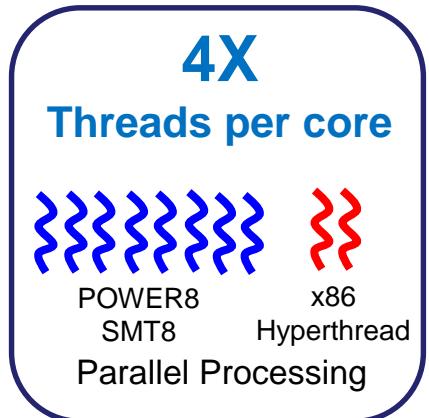
These design decisions result in best performance for data centric workloads like:
Spark, Hadoop, Database, NoSQL, Big Data Analytics, OLTP

SMT=Simultaneous Multi-Threading
OLTP = On-Line Transaction Processing

1. Up to 4X depending on specific x86 and POWER8 servers being compared
2. Up to 6X more cache comparing Intel e7-8890 servers to 12 core POWER8 servers. See speaker notes for more details

IBM®

The OpenPOWER Foundation

Market Shifts	OpenPOWER	New Open Innovation

Market Shifts

- Moore's law no longer satisfies performance gain
- Growing workload demands
- Numerous IT consumption models
- Mature Open software ecosystem

Open Development

open software, open hardware

Collaboration of thought leaders

simultaneous innovation, multiple disciplines

Performance of POWER architecture

amplified capability

New Open Innovation

- Rich software ecosystem
- Spectrum of power servers
- Multiple hardware options
- Derivative POWER chips

The OpenPOWER Foundation is an open ecosystem, using the POWER Architecture to serve the evolving needs of customers.

250+

Members And growing ....

rackspace
the #1 managed cloud company

ubuntu
Supported by Canonical

TYAN

Micron

Mellanox
TECHNOLOGIES

XILINX

I/O
Networking
Storage

Chip
SoC Dev
IP Dev

Technology
FAB

IBM

NVIDIA

SAMSUNG

SYS
ODM
OEM

FW
Open
Source

AVNET

Google

WEB 2.0
Data Center
MSP
Cloud

SW
Linux
ISV
Open Source

Oracle OpenWorld 2017

IBM

IBM Power Systems is open to the core

OpenPOWER
Mellanox TECHNOLOGIES
XILINX
SUPERMICRO
NVIDIA

Open Source Workloads
mongoDB
APACHE Spark
hadoop
EDB POSTGRES
cassandra

OpenCAPI
AMD
Hewlett Packard Enterprise
Google
DELL

Open Frameworks
torch
theano
Chainer
TensorFlow
Caffe

Making Machine Learning and AI more affordable

IBM

IBM Systems and Storage leverage Oracle Software

TS 4500

DS8880

Power E880C

TS 1155

FlashSystem V9000

Storwize V7000F

Power E870C

Enhanced

Power E850C

Power Systems

Enhanced

Power E850C

Enhanced

Power S824

for Business

Linux

AIX

Power S814

Power S822LC HPC

Power S822

POWER8

Power **KVM**   Power **VP**   Power **SC**
Power **HA**   Power **VC**   Power **VM**
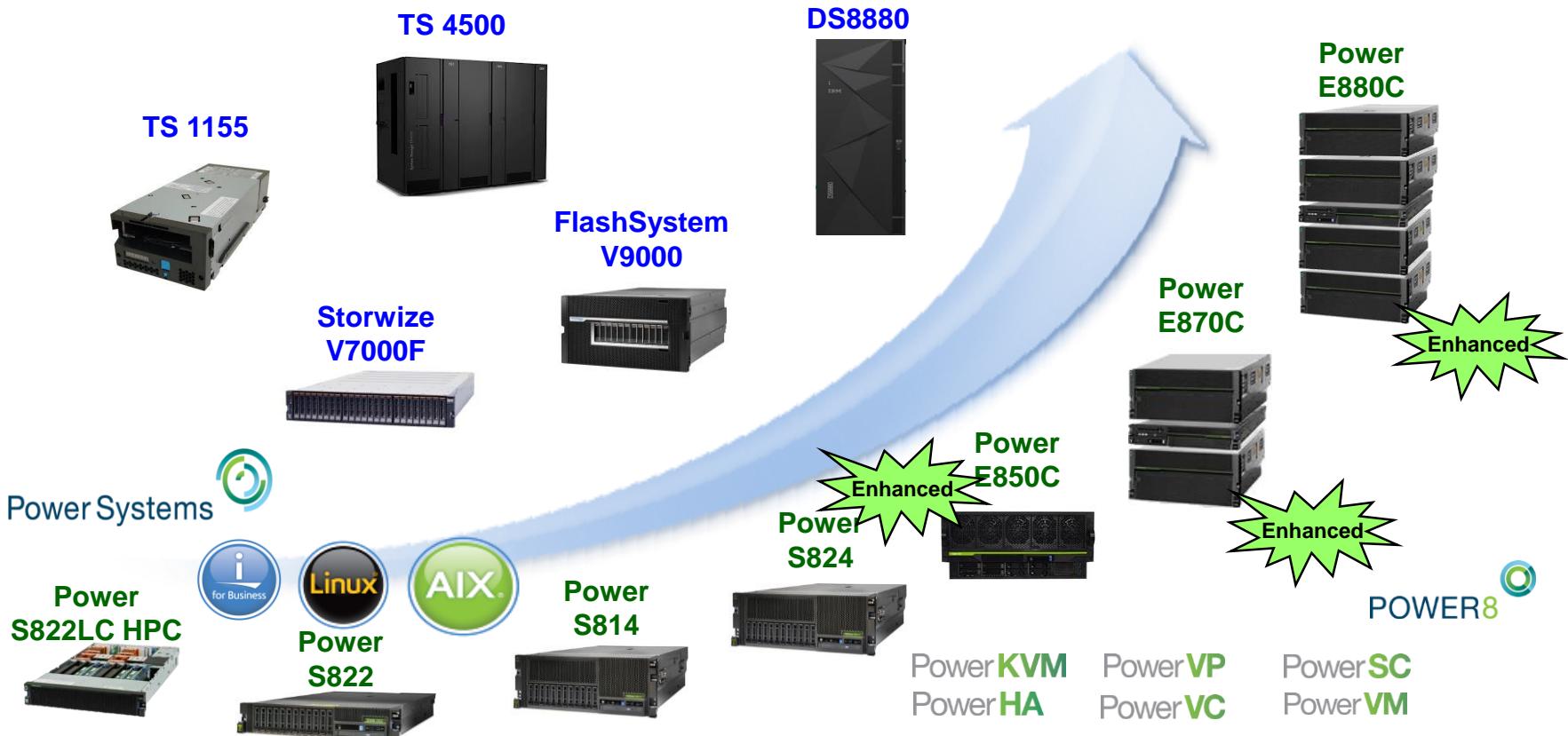
The IBM Power Systems Linux Portfolio

POWER8 is designed for the Big Data era and delivers price-performance leadership to the Linux Market!
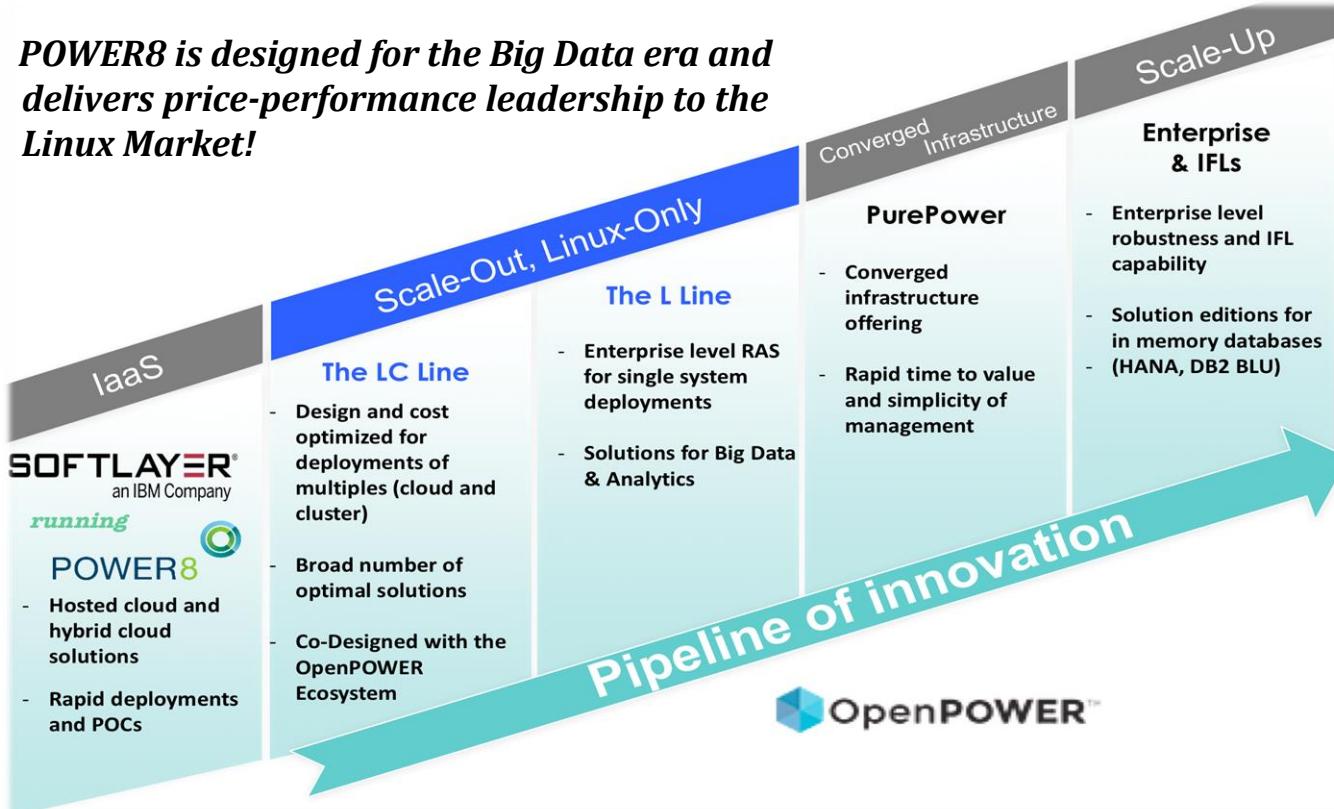
Broad Linux portfolio delivers all your Linux deployment needs

Scale-Up

Converged Infrastructure

Scale-Out, Linux-Only

IaaS

Enterprise & IFLs

- Enterprise level robustness and IFL capability
- Solution editions for in memory databases (HANA, DB2 BLU)

PurePower

- Converged infrastructure offering
- Rapid time to value and simplicity of management

The L Line

- Enterprise level RAS for single system deployments
- Solutions for Big Data & Analytics

The LC Line

- Design and cost optimized for deployments of multiples (cloud and cluster)
- Broad number of optimal solutions
- Co-Designed with the OpenPOWER Ecosystem

SOFTLAYER® an IBM Company
running
POWER8

- Hosted cloud and hybrid cloud solutions
- Rapid deployments and POCs

Pipeline of innovation

OpenPOWER™

IBM Support

redhat   ubuntu
Supported by Canonical

CentOS

Community / 3rd Party Support

debian
fedora   openSUSE

IBM

ML/DL Power Server: S822LC for Big Data and HPC

- Included:

- 2 POWER8 with NVLink

- 4 NVIDIA Tesla P100 GPUs

- 256 GB System Memory

- 2 SSD storage devices

- High-speed interconnect (IB or Ethernet, depending on infrastructure)

- Optional:

- Up to 1 TB System Memory

- PCIe attached NVMe storage

2 Socket, 4GPU System with NVLink
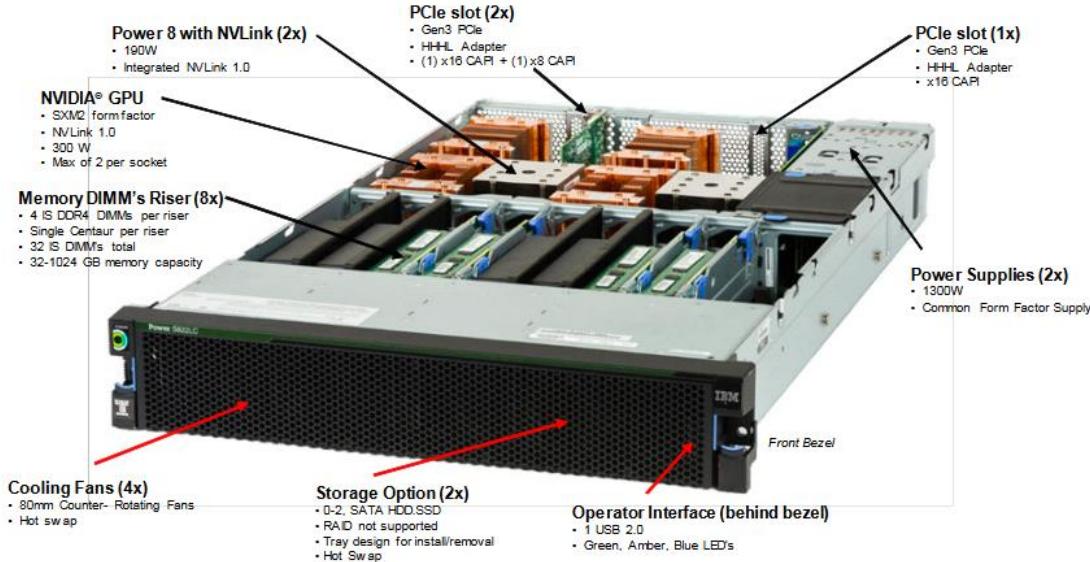
Optimized configuration for Big Data workloads

IBM

IBM Power AI Platform

PowerAI Software Distribution

Deep Learning Frameworks

Caffe	NVIDIA Caffe	IBM Caffe	torch
TensorFlow	theano	Chainer	

Supporting Libraries

DIGITS	OpenBLAS	Distributed Frameworks	Bazel	NCCL

IBM Power System for HPC, with NVLink

Breakthrough performance for GPU accelerated applications, including Deep Learning and Machine Learning.

IBM

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM

The Data Tipping Point that drives Data Lake creation

Oracle OpenWorld 2017

© 2017 IBM Corporation
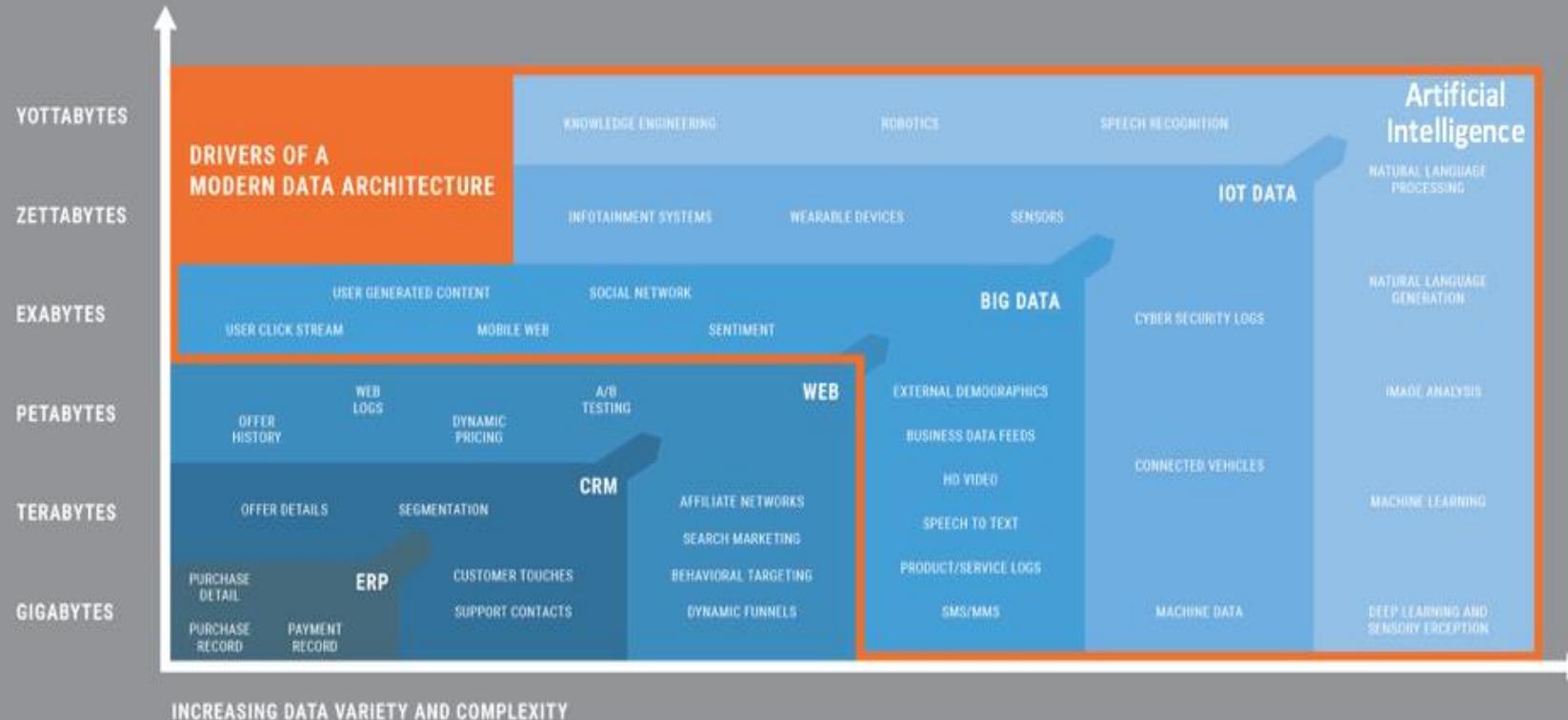
Hortonworks Data Platform, complete on Power Systems

GOVERNANCE INTEGRATION	TOOLS	SECURITY	OPERATIONS

GOVERNANCE INTEGRATION

Data Lifecycle & Governance

Falcon

Atlas

Data workflow

Sqoop

Flume

Kafka

NFS

WebHDFS

TOOLS

Zeppelin   Ambari User Views

DATA ACCESS

Batch	Script	Sql	NoSql	Stream	Search	In-Mem	Others
Map Reduce	Pig	Hive	HBase Accumulo Phoenix	Storm	Solr	Spark	HAWQ Partners Big SQL

Tez   Tez   Slider   S T

YARN: Data Operating System

1

HDFS Hadoop Distributed File System

DATA MANAGEMENT

SECURITY

Administration
Authentication
Authorization
Auditing Data
Protection

Ranger

Knox

Atlas

HDFS Encryption

OPERATIONS

Provisioning, Managing, & Monitoring

Ambari

Cloudbreak

ZooKeeper

Scheduling

Oozie

100% Open Source Connected Data Platform

Oracle OpenWorld 2017

IBM

Hortonworks (HDP) on Power Systems Roadmap

Partnership Announce

Edge Ann Sept 19

Early Availability

HDP 2.5.3 early code on Power Systems

Head start on your Analytics Journey

IBM and Hortonworks assistance for non-prod POCs

General Availability

HDP 2.6 on Power Systems

Proven Power Systems Reference Configs

Hortonworks and IBM production Support

Shared Storage Configs

IBM Spectrum Scale as a HDP certified file system

HDP on Power Systems with Spectrum Scale and Elastic Storage Server

General Availability

HDP 2.6.2 on Power Systems

Simultaneous released with x86

Future Releases

HDP next on Power Systems

Simultaneous releases with x86

HDP is complete on Power Systems

Superior Performance

Maximum Efficiency

Ready for Cognitive

Note: Timing and content subject to change.

3Q2016   4Q2016   Mar 2017   Jun 2017   Aug 2017   later 2017

IBM

IBM and Hortonworks partnered to Deliver:

Data Science Experience (DSX)

Interactive, collaborative, cloud-based environment where data scientists can use multiple tools to activate their insights.

LEARN MORE

Big SQL

Hybrid SQL engine for Hadoop that can connect to disparate sources using a single database connection or query.

LEARN MORE

Power Systems

High performance servers built with open technologies and designed for mission-critical applications

LEARN MORE

Spectrum Scale

Data storage that transforms economics for traditional and next-generation applications with great speed, agility and efficiency.

LEARN MORE

https://hortonworks.com/partner/ibm/

Oracle OpenWorld 2017

IBM®

Flexibility with HDP on Power Systems

- **Scale Up or Out to Meet Evolving Workloads**
  - Scale up each node by exploiting the memory bandwidth and multi-threading
  - 4X threads per core vs x86 allows you to optimize and drive more workload per node
  - 4X memory bandwidth vs x86 gives you more options as your workloads expand and evolve

- **Unmatched Range of Linux Servers**
  - From 1U, 16-core servers up to 16 socket, 192-core powerhouses with industry leading reliability
  - Virtualization options to host low cost dev environments or rich, multi-tenant private clouds
  - Wide range of OpenPOWER servers offered by OpenPower members for on-prem and the cloud

- **Accelerated Analytics**
  - Add accelerators (flash, GPU, FPGA) with direct access to processor memory through OpenCAPI

IBM.

Upgrade your Big Data engine.
Downsize your Big Data server cost.

Get a 3X price-performance advantage for HDP on
IBM Power Systems vs x86 – Guaranteed.

Act Now - Only Available until Dec 31, 2017!

Visit ibm.biz/powerofhortonworks
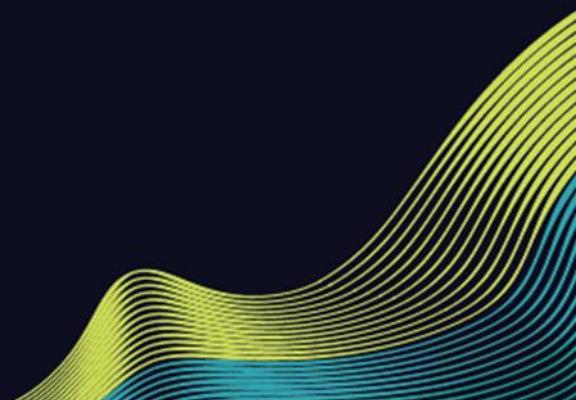
Oracle OpenWorld 2017

Detailed T&C's on page 40

IBM

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM

The connected Data Architecture and Attunity
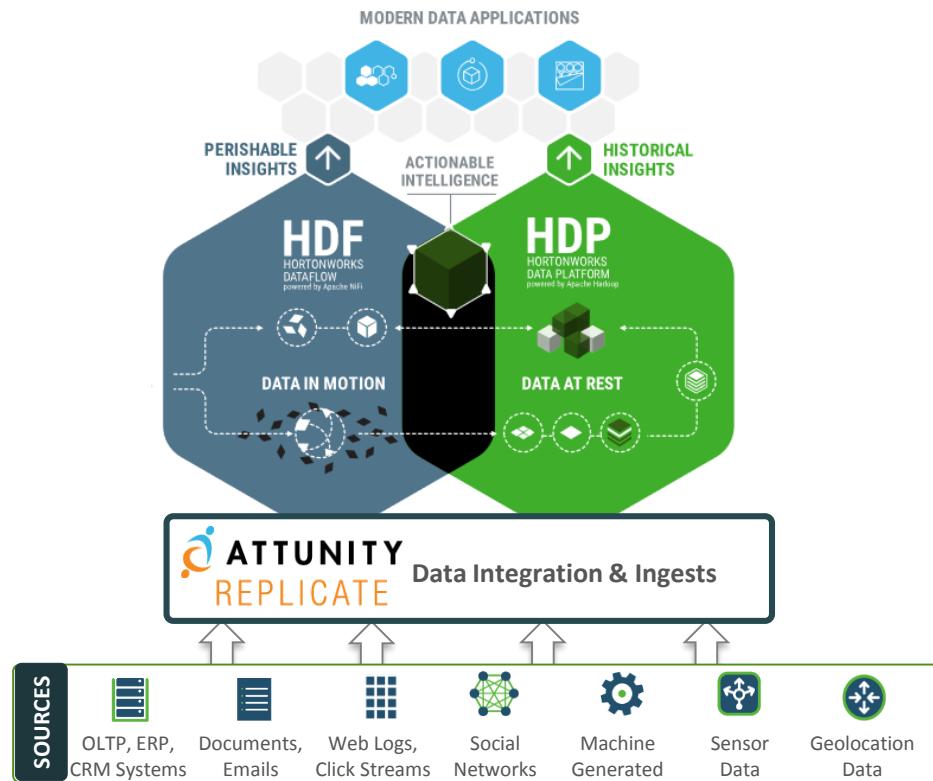
Attunity Replicate for HDP and HDF

Accelerate time-to-insights by delivering solutions faster, with fresher data, from many sources

- Automated data ingest

- Incremental data ingest (CDC)

- Broad support for many sources

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

IBM

Modernize and expand your enterprise data

Data Scientists + Marketing Managers

Enterprise *Data Warehouse*

Big Data Administrators

Off Load

Augment

Explore, Discover, Analyze

Running on POWER8:
The Platform Designed for Big Data

Live *Streams*
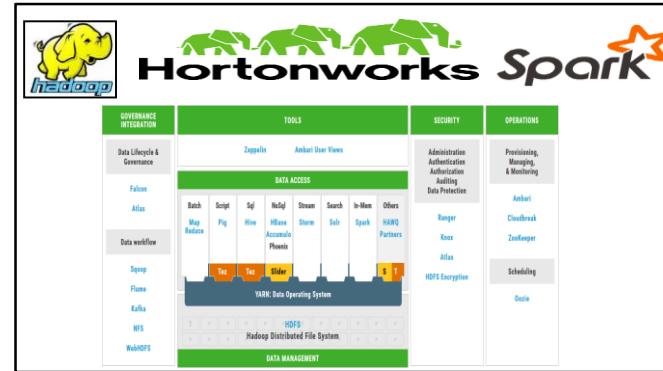
4X threads per core vs. x86

4X memory bandwidth vs. x86

6X more cache vs. x86

Oracle OpenWorld 2017

http://hortonworks.com/solutions/edw-optimization/

IBM

Hadoop Use Cases

- Data Ingestion

- Data Discovery

- EDW Optimization

- Predictive Analytics

- Single View

- And more …

Advertising	Financial Services	Healthcare	Insurance
Manufacturing	Energy & Utilities	Oil & Gas	Pharmaceuticals
Public Sector	Retail	Telecommunications	*Industry Solutions*

IBM

Build Your Data Lake on Linux on Power Systems

- Bulk Load
  - *Oracle Instant Client*
  - *Sqoop*
  - *Attunity Replicate*

- Change Data Capture
  - *Attunity Replicate*

- Federated Queries
  - *Apache Drill*

Extract, Load and Transform data from database sources into Hadoop

Leave (some) data in place

Oracle OpenWorld 2017

IBM

Tools for integrating Oracle and Hadoop on Power Systems are already in place …

Oracle Instant client for Linux on Power connects Hadoop on Linux on Power to Oracle Databases –

Used to fetch Oracle data from Hadoop or issue a Federated Query from Apache Drill

AIX on POWER

Linux on Power
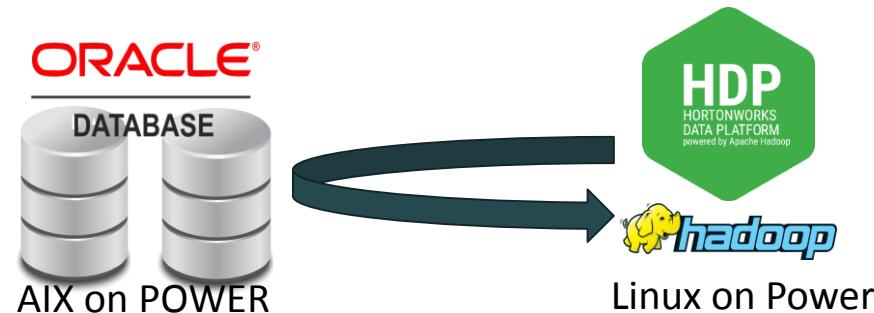
Download the Oracle client here:

http://www.oracle.com/technetwork/topics/linux-power-le-2835260.html

IBM

Tools for integrating Oracle and Hadoop on Power Systems are already in place …

Bulk load with Sqoop:

- Copy Oracle JDBC driver into sqoop's lib directory

```
# ls /usr/hdp/2.6.0.0-598/sqoop/lib/ojdbc6.jar
                /usr/hdp/2.6.0.0-598/sqoop/lib/ojdbc6.jar
```

- Import data from Oracle database into HDFS

```
# /usr/bin/sqoop import --connect jdbc:oracle:thin:system/system@129.40.76.28:1521/ohdppdb.pbm.ihost.com    --username SOE -P
--table SOE.ADDRESSES
… … …
17/09/08 03:29:20 INFO mapreduce.ImportJobBase: Retrieved 913156293 records.
```

- Data Files in HDFS

```
$ hdfs dfs -ls
Found 3 items
drwx------   - hdfs hdfs          0 2017-09-11 12:12 .Trash
drwx------   - hdfs hdfs          0 2017-09-08 03:29 .staging
drwxr-xr-x   - hdfs hdfs          0 2017-09-08 03:29 SOE.ADDRESSES
```

- Import and Create database SOE in Hive similar to source database schema

```
$ sqoop import --connect jdbc:oracle:thin:system/system@129.40.76.28:1521/ohdppdb.pbm.ihost.com --username SOE -P --table
SOE.PRODUCT_DESCRIPTIONS --hive-import
Logging initialized using configuration in jar:file:/usr/hdp/2.6.0.0-598/hive/lib/hive-common-1.2.1000.2.6.0.0-598.jar!/hive-
log4j.properties
OK
Time taken: 3.537 seconds
Loading data to table soe.product_descriptions
Table soe.product_descriptions stats: [numFiles=4, numRows=0, totalSize=111808, rawDataSize=0]
OK
Time taken: 1.058 seconds
```

IBM.

Tools for integrating Oracle and Hadoop on Power Systems are already in place …

Access Oracle data from Spark:

- On Hadoop node running spark, login as spark user and start the spark shell

```
# su – spark
$ cd /usr/hdp/2.6.0.0-598/spark/bin/
$ ./spark-shell
```

- To load the data

```
scala> val x =
sqlContext.read.format("jdbc").option("driver","oracle.jdbc.driver.OracleDriver").
option("url","jdbc:oracle:thin:soe/soe@129.40.76.28:1521/ohdppdb.pbm.ihost.com").o
ption("dbtable", "soe.inventories").load();
x: org.apache.spark.sql.DataFrame = [PRODUCT_ID: decimal(6,0), WAREHOUSE_ID:
decimal(6,0), QUANTITY_ON_HAND: decimal(8,0)]
scala> x.count()
… … …
res1: Long = 901478
```

IBM.

Tools for integrating Oracle and Hadoop on Power Systems are already in place ...

Apache Drill for federated queries:

- Download apache drill from https://drill.apache.org/

- Oracle JDBC driver needs to be placed in /root/apache-drill-1.11.0/jars/3rdparty/ojdbc6.jar

- Modify `drill-override.conf` to include nodes in your cluster. Start apache drill on all the nodes
  ```
  drill.exec: {
  cluster-id: "cluster_id",
  zk.connect: "zookeeper path"
  drill.exec.sys.store.provider.local.path="ojdbc6.jar"
  }
  ```

- Apache drill console could be accessed by http://hostname:8047/storage

- SQL query to Oracle and Hive
  ```
  select * from ORCL_DB.SOE.inventories os, NW_HIVE.inventories hv where
        os.warehouse_id=hv.warehouse_id and os.warehouse_id=221 limit 10;
  ```

IBM.

Tools for integrating Oracle and Hadoop on Power Systems are already in place ...

Create storage plugin for Oracle and Hive

```json
{
  "type": "jdbc",
  "driver": "oracle.jdbc.OracleDriver",
  "url": "jdbc:oracle:thin:soe/soe@hostname:1521/SID",
  "username": null,
  "password": null,
  "enabled": true
}
```

```json
{
  "type": "hive",
  "enabled": true,
  "configProps": {
    "hive.metastore.uris": "thrift://hostname:9083",
    "javax.jdo.option.ConnectionURL":
    "jdbc:mysql://hostname/hive",
    "hive.metastore.warehouse.dir": "/tmp/drill_hive_wh",
    "fs.default.name": "hdfs://hostname:8020",
    "hive.metastore.sasl.enabled": "false"
  }
}
```

Apache Drill   Query   Profiles   Storage   Metrics   Threads   Logs

Sample SQL query: SELECT * FROM cp.`employee.json` LIMIT 20

Query Type

- ● SQL
- ○ PHYSICAL
- ○ LOGICAL

Query

```sql
select * from ORCL_DB.SOE.inventories os, NW_HIVE.inventories hv where os.warehouse_id=hv.warehouse_id and os.warehouse_id=221 limit 10;
```

Submit

Oracle OpenWorld 2017

© 2017 IBM Corporation

IBM

Change Data Capture for many sources/targets:
Attunity Replicate, Hortonworks HDP, IBM Power Systems

Automated data ingest

Process data in real time

Rapid deployments of <u>HUGE</u> data lakes

Incremental updates with CDC

Continuous data refresh for <u>RELEVANT</u> analytics

Broad support for many enterprise data sources

<u>COMPLETE</u> datasets across DBs, DWs & mainframes

HORTONWORKS
PARTNERWORKS
HDP CERTIFIED
SEC READY

HORTONWORKS
PARTNERWORKS
HDP CERTIFIED
READY

HORTONWORKS
PARTNERWORKS
HDP CERTIFIED
YARN READY

ATTUNITY

Hortonworks
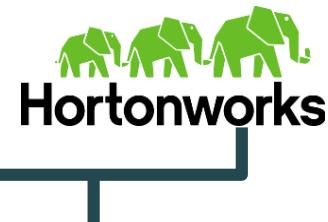
Data Sources Include
DB2, Hadoop, SQL Server, MySQL/MariaDB,
Oracle, PostgreSQL, SAP Sybase ASE,
Teradata Database, ODBC, Amazon RDS for
Aurora, MySQL, Oracle

Running on POWER8:
The Platform Designed for Big Data

4X
threads per core

4X
memory bandwidth

6X
more cache

IBM.

Experiences: Bring Oracle data into Hive using Attunity Replicate

AIX on POWER

ATTUNITY REPLICATE
Currently X86

Linux on Power

IBM
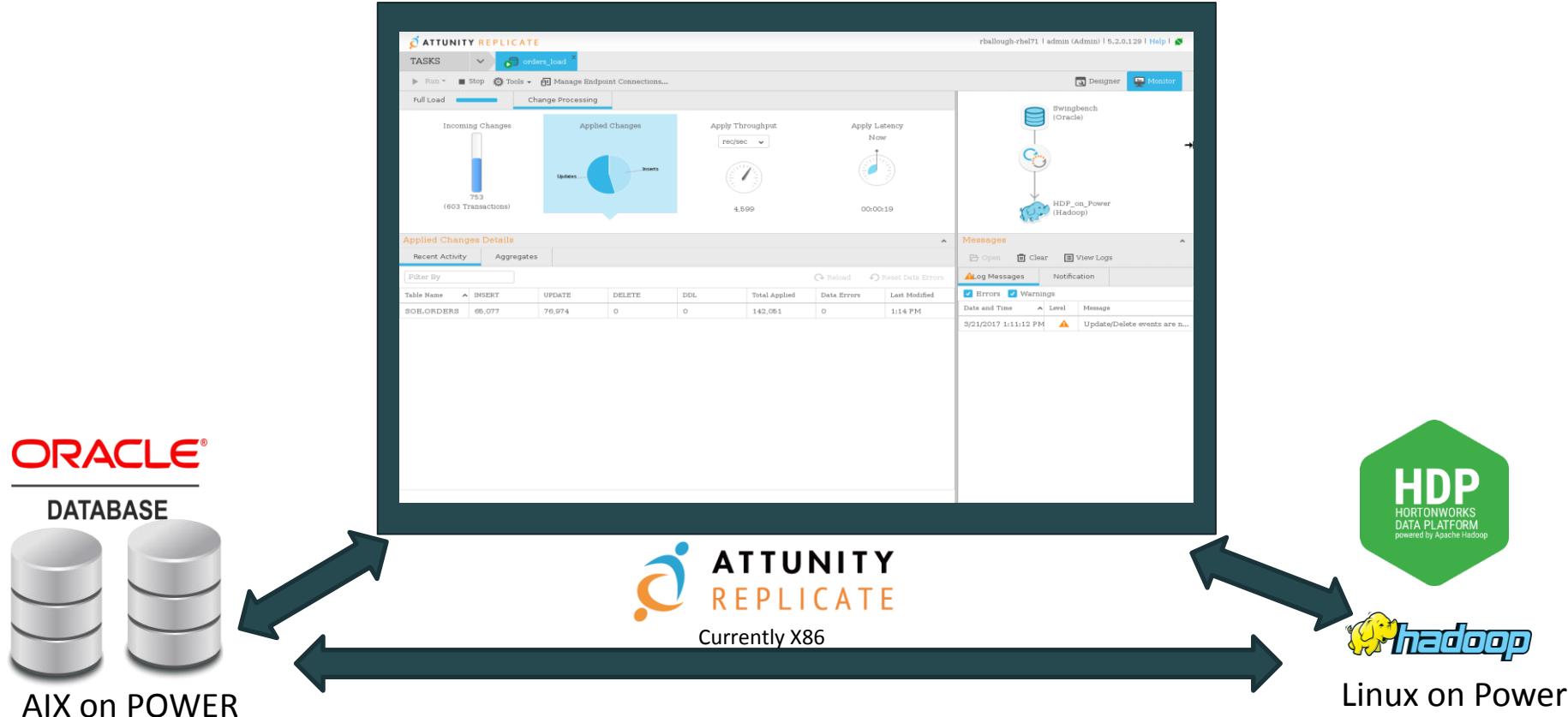
Configuration Details

Oracle Database

- *Version 12cR2*
- *AIX 7.2 on POWER8*
- *ARCHIVELOG Mode is On*
- *Supplemental Logging is enabled*
- *Basic encryption,OLTP and HCC supported*
- *Required privileges to be granted be V$xxx*

Attunity Replicate

- *Version 5.5.0.143*
- *RHEL 7.2*
- *Oracle Instant Client for Linux (x86-64)*
- *Minimum of 8 GB RAM, and 320 GB of disk space*
- *Mozilla Firefox Version 38 and above,*
- *Minimum Network 1 Gb interface*

Horotnworks HDP

- *Version 2.6.1*
- *RHEL 7.2 on POWER8*
- *Hadoop WebHDFS must be accessible*
- *Hadoop WebHCat service must be running*
- *Hadoop target settings must have access to Hcatalog*
- *Access Hive using ODBC 2.2 and above*

IBM

Attunity Replicate: Manage Endpoint – Oracle *as source*

Oracle OpenWorld 2017

© 2017 IBM Corporation

IBM®
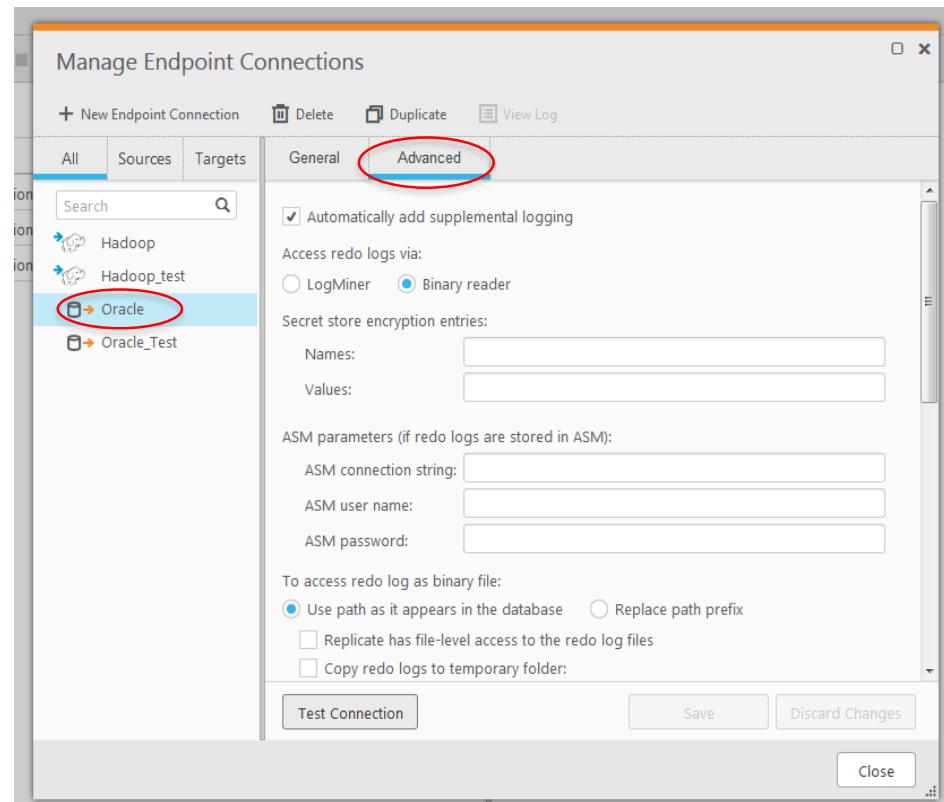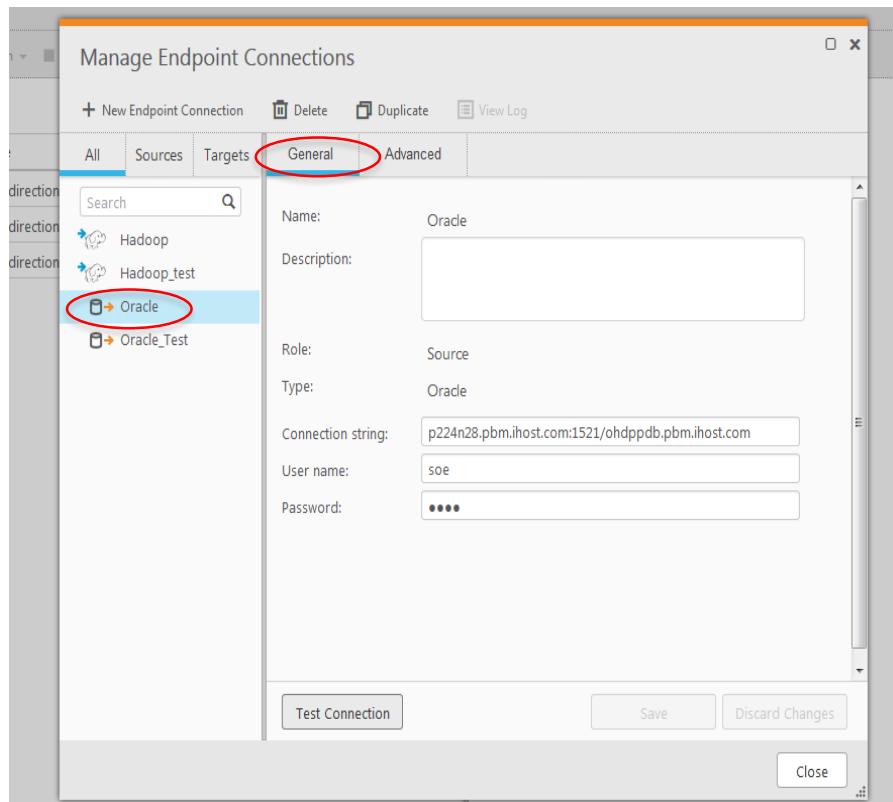
Attunity Replicate: Manage Endpoint – Hadoop *as target*

Oracle OpenWorld 2017

Attunity Replicate: Task (Full Load)

Oracle OpenWorld 2017

Attunity Replicate: Task (Incremental Change)

Oracle OpenWorld 2017

Agenda

IBM and Oracle Collaboration

IBM Power Systems

Hortonworks Data Platform

Attunity Replicate

Experiences Ingesting Oracle Data into Hadoop on Power Systems

Conclusion and Acknowledgments

Live Demonstration Presented @
IBM Booth, Systems Pedestal

IBM.

Only Available until Dec 31, 2017!
Hortonworks HDP 3X POWER8 Price-Performance Guarantee

IBM Power Systems guarantees the Power S822LC for Big Data system built with POWER8 delivers at least a 3X price-performance advantage vs. x86 based results when running a customer application/workload with Tez/Hive LLAP on Hortonworks HDP under the conditions noted below. A Worker Node is a server carrying out the HDP query functions, with one Worker Node per server.

3X price-performance means that the customer's documented throughput performance on the cluster of S822LC for Big Data Worker Nodes divided by the price of the cluster of Worker Nodes will be at least 3 times higher than the customer's documented throughput performance on the cluster of x86 based Worker Nodes divided by the price of the cluster of x86 Worker Nodes.

EX: If queries per second on the cluster of S822LC Worker Nodes are 30,000 and 10,000 on the cluster of x86 based Worker Nodes, while the price of the S822LC Worker Node cluster is $10,000, and the price of the x86 based Worker Node cluster is $10,000, then the Throughput Performance Per Price would be exactly 3 times higher and the guarantee would be met."

The IBM Power S822LC for Big Data servers (22-core/2.89 GHz) used as Worker Nodes must be purchased from IBM or an authorized IBM Business Partner prior to September 30, 2017. The guarantee period is valid for three (3) months from the date of purchase. The x86-based Worker Nodes must be comparably configured branded servers from Cisco, Dell, HP, or Lenovo and the client is responsible for all Hortonworks licenses.

3X throughput performance per price means that the customer's documented throughput performance on the cluster of Power S822LC for BD Worker Nodes based on either queries, operations or transactions per second divided by the price of the cluster of Worker Nodes will be at least 3 times higher than the customer's same documented throughput performance on the cluster of x86 Worker Nodes divided by the price of said cluster of x86 Worker Nodes.

Remediation: IBM will provide additional performance optimization and tuning services consistent with IBM Best Practices, at no charge. If unable to reach the guaranteed level of price-performance, IBM will provide additional equally configured Worker Nodes to those already purchased to reach the guaranteed level of price-performance.

Notes:
1. Client's Power S822LC for BD Worker Nodes and the x86 Worker Nodes must be running at similar utilization rates of at least 50% or higher, using the same software stack as described in Note #4, and which are configured similarly.
2. Client's Power S822LC for BD performance cannot be constrained by I/O subsystem. Specifically, the I/O subsystem on the Power S822LC for BD Worker Node must achieve greater than or equal I/O bandwidth and operations per second than the x86 Worker Node.
3. Client's Power S822LC for BD Worker Node's physical memory must be the same or greater than the physical memory on the x86 Worker Node.
4. Applicable software stack is Tez/Hive LLAP on HDP 2.6 or later for both the Power S822LC and x86-based Worker Nodes.
5. Client is responsible for demonstrating comparable real-world representative workload between the Power S822LC for BD Worker Node and the x86 Worker Node through the use of the IBM provided tools and comparable tools on x86 systems.
6. 3X guarantee is based on a list price for x86 servers from Dell, Cisco, HP or Lenovo based on E5-2600 v4 or earlier processor technology and the IBM S822LC for Big Data.

IBM

Acknowledgments

The authors graciously thank:

Mithun Hr for his full support and execution of the runs that supported this presentation.