BigData & Analytics
Linux on POWER
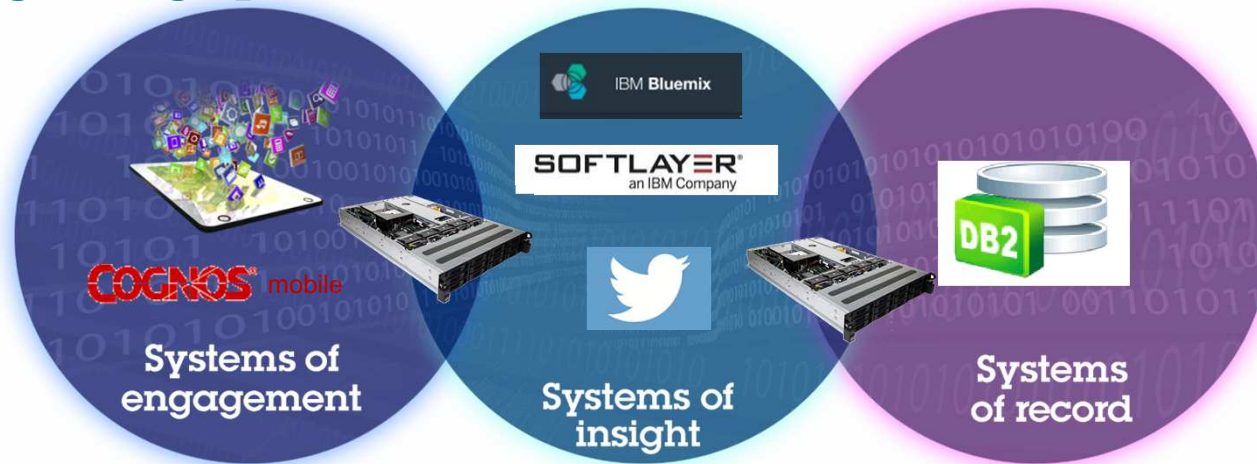
Angel González

Analytics Client Architect
angelito@de.ibm.com

NextGen Analytics Systems
emerge as SOI growing upon SOE and SOR

IBM Bluemix

SOFTLAYER
an IBM Company

COGNOS mobile

DB2

Systems of engagement

Systems of insight
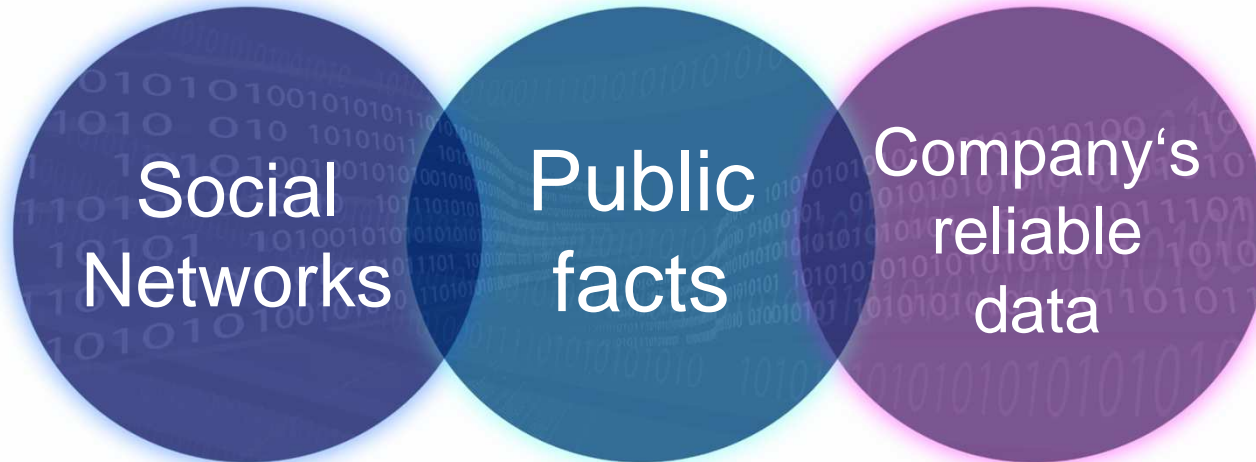
Systems of record

Mobile users

Data analysts

OLTP transactions placing orders

NextGen Analytics Systems

exploit all information sources, no matter of their nature, format, origin, etc.

Social Networks

Public facts

Company's reliable data

- Sentiment
- Consumer rating
- Celebrity trends

The Weather Channel

eventful   Location: Madrid metro area

Events   Movies   Demand it!   My Eventful

- Events
- Weather
- Traffic

- Transactions
- Master Data
- Legacy DWH
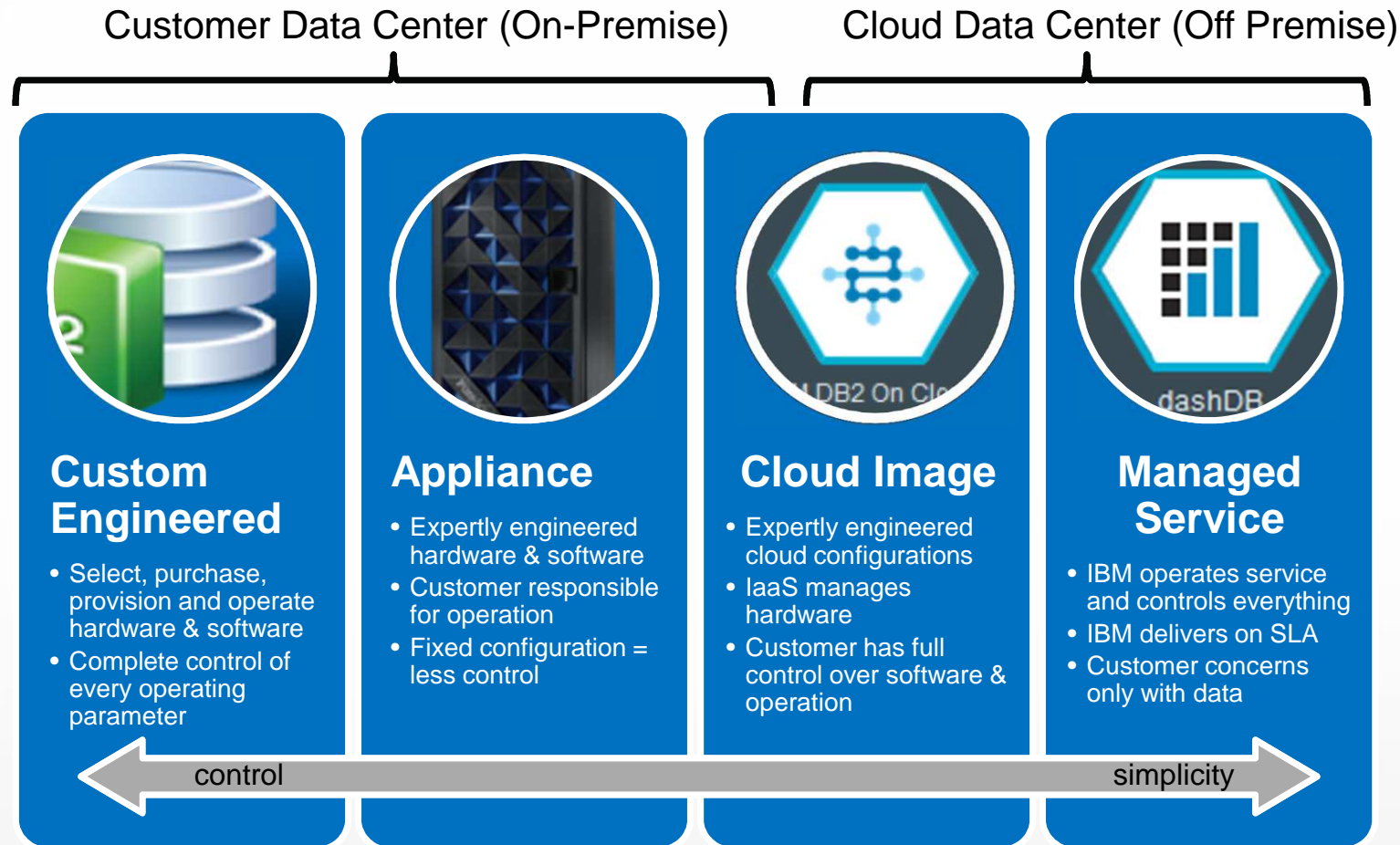
NextGen Analytics Systems
adopt innovative technological and architectural patterns like HTAP

https://www.gartner.com/doc/2657815/hybrid-transactionanalytical-processing-foster-opportunities

Gartner Research Note G00259033: Gartner 01-2014 Hybrid Transaction Analytical Processing Will Foster Opportunities

High Volumes of Data in Memory

Maximum Memory bandwidth

	Haswell EX	POWER8
SMT options	1,2*	1, 2, 4, 8
Max L1 Cache	64 KB	96KB
Max L2 Cache	256 KB	512 KB
Max L3 Cache	45 MB	96 MB
Max L4 Cache	0	128 MB
Memory Bandwidth	102 GB/s	230 - 410 GB/s

Memory bandwidth sustained GB/s
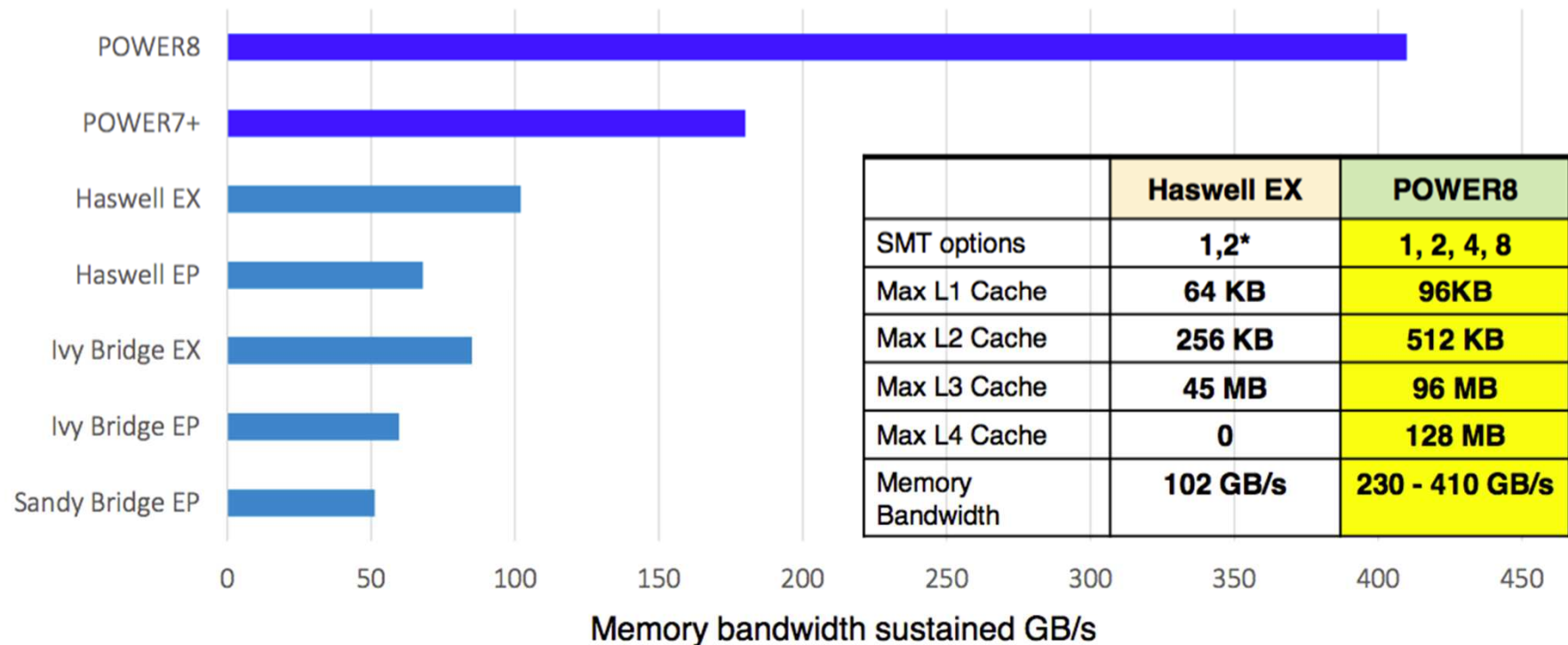
Little Endian DB2 with BLU Acceleration on POWER8 delivers 1.90X more query results per hour per core than Intel Haswell

Gain insights faster with the same software on POWER8 versus Intel

- Deliver 1.90X more query results per hour per core running Little Endian DB2 with BLU Acceleration on POWER8 versus Intel Haswell (E5-2699 v3)

- Based on 100 users concurrently executing 100 distinct queries each

POWER8
IBM S824
24c/192t
DB2 / Ubuntu

x86
HP DL380p
E5-2699 v3
36c/72t
DB2 / RHEL
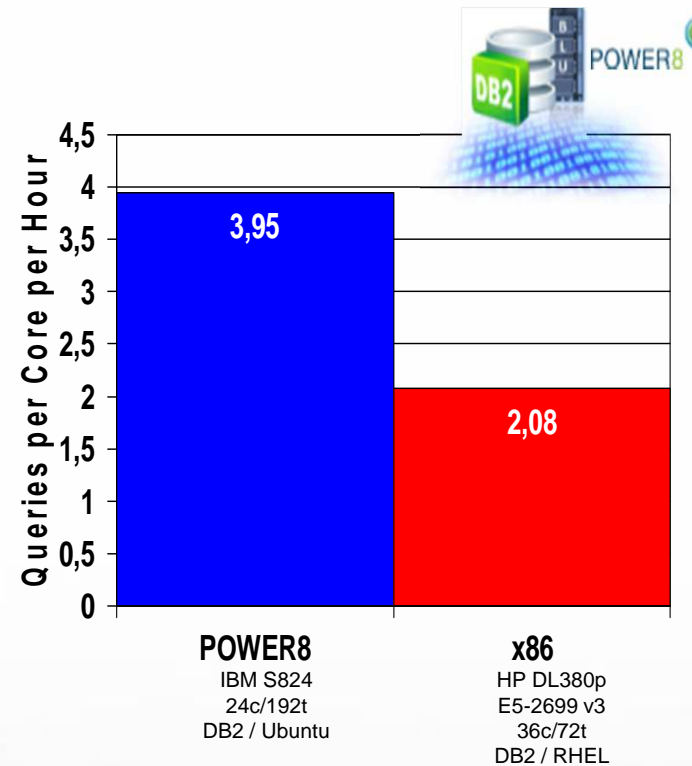
- 1.90X more query results is based on IBM internal testing of a sample analytic workload; current as of April 24, 2015. Performance improvement figures are cumulative of all queries in the workload. Individual results will vary depending on individual workloads, configurations and conditions.
- IBM Power System S824; 24 cores / 192 threads, POWER8; 3.5GHz, 256 GB memory, DB2 10.5 / Ubuntu 14.04
- Competitive stack: HP DL380p; 36 cores / 72 threads; Intel E5-2699 v3; 2.3 GHz; 256 GB; DB2 10.5 / RHEL 7.1

Most of your IT costs lay beneath the surface

*"For modern IT platforms, **30% of the total cost** is the cost of acquiring the **equipment**. The balance is for IT labor/services to configure, maintain, upgrade, reconfigure, and ultimately decommission the equipment."* [1]

*"IT organizations are spending more than **70% of their total IT** budgets on **maintenance and ongoing operations**."* [2]

10%-50%

Acquisition Cost

Deployment

Maintenance

Downtime

Security

Software

Asset Management

Power

Space

50%-90%

TCO

[1] IDC – "IT Capital Investments: Evaluating Technology Life-Cycle Management and Lease-Versus-Own Options"
[2] National Analysts – "IBM Market Intelligence Time to Value Study"

Most of your IT costs lay beneath the surface

IBM POWER8 S824L System --PowerVM

24 core POWER8
512GB RAM
HANA on POWER
SLES 11 SP3, XFS
Flash System 840

PowerVM

3YR TCA
$180,914
Excluding HANA Software

x86 Haswell Server – Competitive Hypervisor v5.5

24 core Haswell Xeon E5-2680 v3
512 GB RAM
HANA on x86
SLES 11 SP3, XFS
Flash System 840

Competitive Hypervisor v5.5

3YR TCA
$159,084
Excluding HANA Software

1TB (uncompressed) Operational Analytics workload

434	Simple Reports/Hour
39	Intermediate Reports/Hour
49	Complex Reports/Hour

✓ 2.51x Faster Simple Reports
✓ 1.61x Faster Intermediate Reports
✓ 1.13x Faster Complex Reports

173	Simple Reports/Hour
24	Intermediate Reports/Hour
43	Complex Reports/Hour

45%
Better price performance

Based on IBM internal tests (BDInsights) comparing SAP HANA on Power system with a comparably priced, comparably tuned competitor configuration (version available as of 11/1/2015) executing a materially identical 1TB Uncompressed Operational analytics workload in a controlled laboratory environment. Test measured 100 concurrent user report throughput executing identical SQL query workloads. 3YR Total Cost of Acquisition (TCA) based on publicly available U.S. prices current as of Nov 1, 2015, including hardware, software, and maintenance. Compared prices exclude applicable taxes, and are subject to change without notice. Competitor configuration: Haswell EP 2s/24c, 2.5 GHz, xeon E5-2680 v3, 512GB RAM, SLES 11SP3, VMware v5.5, XFS and IBM FlashSystem v840 running SAP HANA for x86. IBM configuration: POWER8 S824L 2s/24c, 3.52 GHz, 512GB RAM, SLES 11SP3, PowerVM, XFS and IBM FlashSystem v840 running SAP HANA for Power Systems. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment.

IBM POWER Analytics Solutions

Analytics solutions

Unlock the value of data with an IT infrastructure that provides speed and availability to deliver accelerated insights to the people and processes that need them.

IBM Data Engine for Analytics - Power Systems Edition

A customized infrastructure solution with integrated software optimized for both big data and analytics workloads.

IBM Data Engine for NoSQL – Power Systems Edition

Unique technology from IBM delivers dramatic reductions in the cost of large NoSQL databases.

SAP HANA benefits from the enterprise capabilities of Power Systems

SAP HANA runs on all POWER8 servers. Power Systems Solution Editions for SAP HANA BW are easy to order and tailored for quick deployment and rapid-time-to value, while offering flexibility to meet individual client demands.

DB2 with BLU Acceleration on Power Systems

Enable faster insights using analytics queries and reports from data stored in any data warehouse, with a dynamic in-memory columnar solution.

IBM Solution for Analytics – Power Systems Edition

This flexible integrated solution for faster insights includes options for business intelligence and predictive analytics with in-memory data warehouse acceleration.

IBM Data Engine for Hadoop and Spark – Power Systems Edition

A fully integrated Hadoop and Spark solution optimized to simplify and accelerate unstructured big data analytics.

http://www-03.ibm.com/systems/power/solutions/bigdata-analytics/index.html
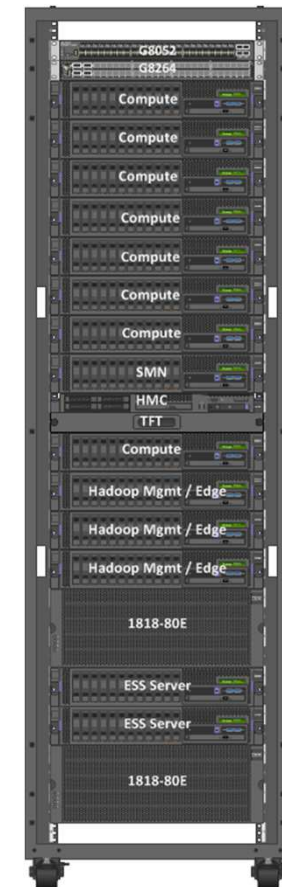
IBM Data Engine for Analytics

A fully integrated solution with software and infrastructure optimized for Big Data & Analytics
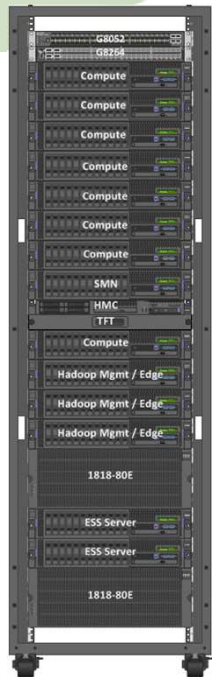
- ✓ **Simplify operations – easy to deploy and manage**
- ✓ **Preloaded with IBM BigInsights and IBM Open Platform**
- ✓ **Designed for mixed analytics workloads: streams, at rest, text**
- ✓ **Enterprise grade Hadoop with advanced resource and storage management**
- ✓ **Adapt and scale to your changing analytics needs**

Appliance-Like but much more Versatile!

Single vendor support

Less than half storage infrastructure with only **1** copy of data*

1.5x to 2.2x better per core performance vs x86 configs**

Lowest $/TB and **over a third more** usable storage***

Built on POWER8: The Platform Designed for Big Data

4X	**4X**	**5X**
threads per core vs. x86 (up to 1536 threads per system)	memory bandwidth vs. x86 (up to 16TB of memory)	more cache vs. x86 (up to 224MB cache per socket)

* Compared with a standard triple replica Hadoop configuration.
** Based on internal Terasort and SparkBench results
** List price vs full rack configuration vs Oracle Big Data Appliance with data connectors and BigSQL

IBM POWER Analytics Solutions

IBM Software > Products > Data management platform >
Hadoop > IBM BigInsights >

IBM Open Platform with Apache Hadoop

100% Apache Hadoop Open Source platform

No-charge download available
→ IBM Open Platform with Apache Hadoop

IBM released a new Open Platform for Apache Hadoop on Intel and Power platforms and IBM BigInsights v4.1 on Intel and Power Systems.

System Management Node - S812LC
- 8 x 3.32 GHz cores
- 32 GB RAM, up to 1 TB
- 2 x 1 TB 3.5" SATA HDD (OS)
- Dual-port 10 GbE (data)
- Dual-port 1 GbE (mgmt)

Hadoop Management Node - S812LC
- 10 x 2.92 GHz cores
- 128 GB RAM, up to 1 TB
- 2 x 1 TB 3.5" SATA HDD (OS)
- Dual-port 10 GbE (data)
- Dual-port 1 GbE (mgmt)

Data Node - S812LC
- 10 x 2.92 GHz cores
- 128 GB RAM, up to 1 TB
- 2 x 1 TB 3.5" SATA HDD (OS)
- 12 x 6 TB 3.5" SATA HDD (data)
- Dual-port 10 GbE (data)
- Dual-port 1 GbE (mgmt)
- PMC Sierra 71605E RAID adapter
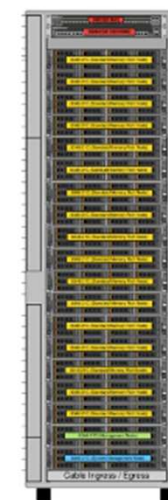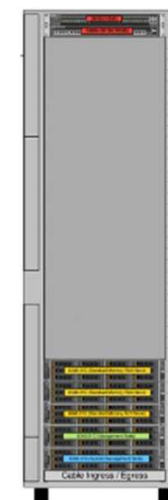
Configuration	Minimum	Maximum
Data Nodes (cores, memory)	3 DN (30 cores, 384 GB, up to 3 TB)	17 DN (170 cores, 2176 GB, up to 17 TB)
Data Nodes (raw / effective* TB, disks)	3 DN (216 / 162TB, 36 disks)	17 DN (1224 / 918 TB, 204 disks)
Data Nodes (links)	3 DN (3 x 10Gb / 6 x 1Gb)	17 DN (17 x 10Gb / 34 x 1Gb)
Hadoop Mgmt (links)	1 MN (1 x 10Gb / 2 x 1Gb)	1 MN (1 x 10Gb / 2 x 1Gb)
System Mgmt Node (links)	1 SMN (1 x 10Gb / 2 x 1Gb)	1 SMN (1 x 10Gb / 2 x 1Gb)
Total 10Gb links, 1Gb links	5 x 10Gb, 10 x 1Gb	19 x 10Gb, 38 x 1Gb
Network Switches	1 x G8264/G8124 + 1 x G8052	1 x G8264/G8124 + 1 x G8052

* As a rule of thumb, reserve 25 percent of the total disk space (raw capacity) for the local file system as shuffle file space

Power Systems

IBM

IBM Open Platform with Apache Hadoop

- 100% open source based on Apache products defined by ODPi
- Flexible platform for processing large volumes of data
  - Includes Apache Hadoop, Ambari, Spark and more
- Install just the components you want
- Requires RHEL 7.1
- Download from http://www-03.ibm.com/software/products/en/ibm-open-platform-with-apache-hadoop (Spark 1.5.1)

IBM Open Platform with Apache Hadoop

HDFS	MapReduce2	Spark	Hive	HCatalog	Pig
YARN	Ambari	HBase	Flume	Sqoop	Solr/Lucene

Apache Open Source Components
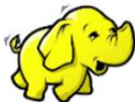
IBM BigInsights Enhancements

Choose from 3 Value Add Services
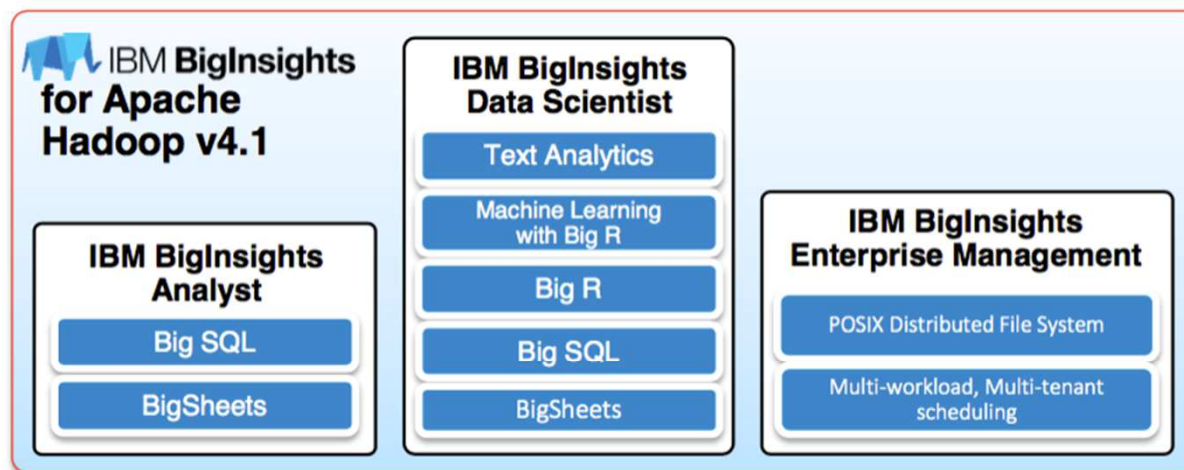
Free 100% Open Source

More info:
https://developer.ibm.com/hadoop/docs/getting-started/faqs/

Why BigInsights?

Big SQL is the
only engine that can
execute all 99 queries
from the Hadoop-DS
benchmark with minimal
porting effort
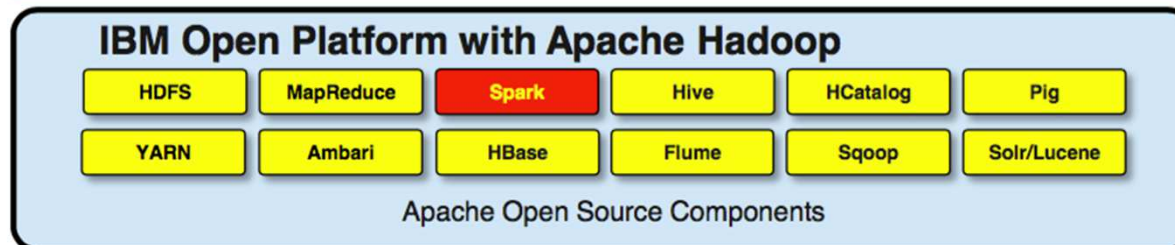
| Runs without Modification |
| Runs with Minor Modification |
| Extensive Modification |
| Not Working |

Porting Effort: 1 hour | 3-4 weeks | Indefinite

IBM Power Systems S812LC
optimized for entry and small Hadoop workloads

Improve the agility and reduce the cost of running Spark and Hadoop workloads

- 16X the memory capacity of Xeon E3 servers, 2X 1P Xeon E5 servers
- Complete the same Spark workloads for <½ the cost of Intel Xeon E5-2690 v3 systems
- 2.3X BETTER performance per dollar spent
- 94% more Spark workloads in the same rack space as Intel Xeon E5-2690 v3 systems
- 1.94X BETTER performance per system (10 core S812LC vs 24 core DL380)

- 1-socket, 2U
- Up to 10 cores (2.9-3.3Ghz)
- 1 TB Memory (32 DIMMs)
- 115GB/sec memory bandwidth
- 14 LFF (HDD/SSD) 84TB storage
- 4 PCIe slots, 2 CAPI enabled
- Default 3 year 9x5 warranty, 100% CRU

Power S812LC – Designed for BigData
Optimized price-performance for the unique needs of Spark

Designed for the Cognitive Era to Make Better Decisions even Faster

- IBM Data Engine for Hadoop and Spark infrastructure delivers Spark workload scaling to minimize execution times and reduce batch windows

  - **2.1X** more performance per dollar spent for Spark Logistic Regression based **Machine Learning** used in **model training** by wide variety of lines of business

  - **1.4X** more performance per dollar spent for Support Vector Machine (SVM) – a **Machine Learning** algorithm used in product **Recommender** Systems

  - **1.7X** more performance per dollar spent for **Spark SQL query processing** used widely in Big Data clusters

- All results are based on IBM Internal Testing of 3 SparkBench benchmarks consisting of SQL RDD Relation, Logistic Regression, SVM
- 6 Data Nodes and 1 Management Node. Each node is IBM Power System S812LC 10 cores / 80 threads, POWER8; 2.92GHz, 256 GB memory, RedHat 7.2, Spark 1.5.1, OpenJDK 1.8
- 6 Data Nodes and 1 Management Node. Each node is x86 E5-2620V3 12 cores / 24 threads, E5-2620 V3; 2.4GHz, 256 GB memory, RedHat 7.1, Spark 1.5.1, OpenJDK 1.8
- Pricing is based on web prices of HP DL380 and list prices of IBM Power S812LC

Logistic Regression, SQL, and SVM Performance

Logistic Regression, SQL, and SVM Price-Performance

Power Systems
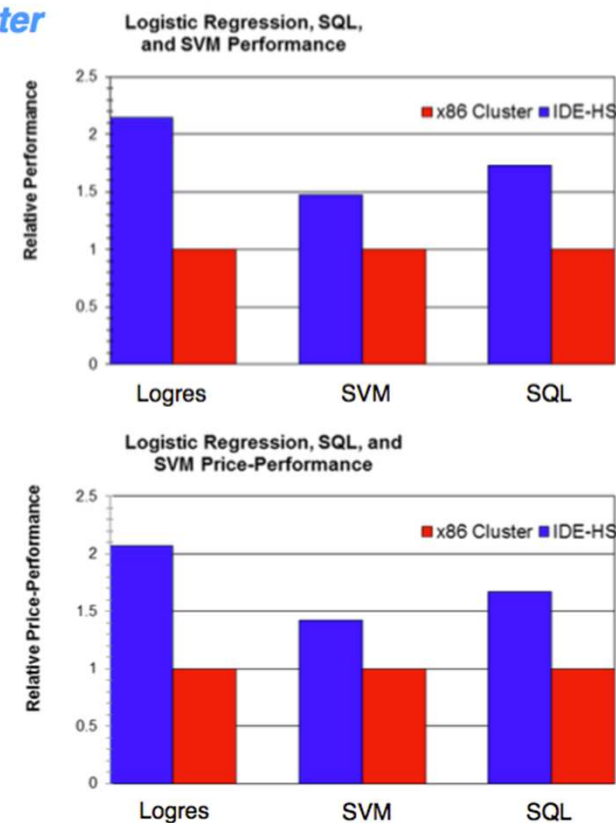
IBM Systems (6)	IBM Security (23)	IBM Analytics (53)	

IBM Analytics (53) — left column:

- DB2 Advanced Enterprise Server Edition (w/BLU)
- DB2 Advanced Workgroup Server Edition (w/BLU)
- DB2 Connect Application Server Advanced Edition
- DB2 Connect Application Server Edition
- DB2 Connect Enterprise Edition
- DB2 Connect Unlimited Advanced Edition for System Z
- DB2 Connect Unlimited Edition for System i
- DB2 Connect Unlimited Edition for System z
- DB2 Developer Edition
- DB2 Enterprise Server Edition
- DB2 Express Edition
- DB2 Merge Backup for Linux, UNIX, and Windows
- DB2 Recovery Expert for Linux, UNIX, and Windows
- DB2 Workgroup Server Edition
- IBM Analytical Decision Management
- IBM BigInsights
- IBM BigMatch
- IBM Cognos Analytics (Formerly IBM Cognos Business Intelligence)
- IBM Data Server Driver for JDBC and SQLJ
- IBM Data Server Driver for ODBC and CLI
- IBM Data Server Driver Package
- IBM Data Server Manager - Base Edition
- IBM Data Server Manager - Base Edition (Continuous Delivery)
- IBM Data Server Manager - Enterprise Edition
- IBM Data Server Manager - Enterprise Edition (Continuous Delivery)
- IBM DB2 BLU Acceleration In-Memory Offering
- IBM DB2 Configuration Manager for z/OS
- IBM DB2 Encryption Offering
- IBM DB2 Performance Management Offering
- IBM DB2 Query Workload Tuner for z/OS
- IBM Informix Server
- IBM InfoSphere Identity Insight
- IBM InfoSphere Master Data Management
- IBM InfoSphere Optim Performance Manager Extended Edition for DB2 for Linux, UNIX, and Windows
- IBM InfoSphere Optim Performance Manager Extended Insight
- IBM InfoSphere Optim Performance Manager for DB2 on Linux, UNIX, and Windows
- IBM InfoSphere Optim pureQuery Runtime for Linux, UNIX and Windows
- IBM InfoSphere Optim Query Workload Tuner for DB2 for LUW
- IBM Netezza NPS
- IBM SPSS Analytic Server
- IBM SPSS Collaboration and Deployment Services
- IBM SPSS Data Access Pack
- IBM SPSS Modeler
- IBM SPSS Modeler Gold
- IBM SPSS Modeler Server
- IBM SPSS Statistics

IBM Analytics (53) — right column:

- IBM SPSS Statistics Server
- InfoSphere Data Replication
- InfoSphere Streams
- Optim High Performance Unload for DB2 for Linux, UNIX, and Windows
- IBM Data Studio (ManageTo)
- IBM DB2 Query Management Facility for z/OS (ManageTo)
- InfoSphere Data Architect (ManageTo)

IBM Watson (1)

IBM Industry Solutions (1)

IBM Watson IoT (47)

IBM Cloud (42)

IBM Systems: Sys SW + Storage (19)

IBM Commerce (4)

18

IBM Power Open Source Ecosystem

Dev. Env /Tools

Databases & search engines

Big Data & Analytics

Cloud Managemnt Stack

Other HA, Security etc.

Technical Computing

Available:
Angular.js, Backbone, Bootstrap, buildbot, busyBox, Continuum, CruiseControl, Docker , Eigen lib, Erlang, Ganglia, GCC, gccgo, GDB, GoLang, Gump, Jenkins, Jruby, Lynx, LLVM, logstash, logstash-forwarder, kibana, Maven, Nagios, NGINX, node.js, OCaml, OpenJDK, Phantom.js, PHP, phpMy Admin, Perl, PyCi, Python, Python-Django, Python-Pip ecosystem, PyPy, R, RabitMQ, rsyslog, Ruby, Ruby on Rails (rbenv), Ruby Gems, scala, snappy, Socket.io (npmjs), SpiderMonkey, Strider, Supervisord, SystemTap, Vagrant, V8, Xerces

Optimized:
R

Port In Progress:
Travis-CI

Available:
Accumulo (column), Cassandra
CouchDB (document)
Couchbase (noSQL)
Derby, InfluxDB, Firebird
ElasticSearch
EnterpriseDB, LevelDB, MariaDB (v10.x optimized)
Memcached (KVS)
MongoDB 3.2 (document)
MySQL, Percona Server, PostgreSQL, PrestoDB, RabbitMQ, Redis (KVS), Riak, Sphinx, SQLite, TokyoCabinet
Virtuoso (graph)
Wired Tiger, ZODB

Port In Progress:
GPUDB, Neo4j (graph)

Optimizing:
CouchDB

Evaluating:
InfiniSQL

Available:
Hadoop Base, Hive, HBase, Accumulo, Ambari, Avro, ElasticSearch
Falcon, FluentD
Flume, Hue, Kafka, Knox, Lucene-Solr, Mahout, Oozie, Parquet, Phoenix
Pig, Ranger, Slider, Spark, Sqoop, Storm, Tez, Zookeeper

Optimizing:
Hadoop

Evaluating:
Clusterpoint

Available: Open source application is ported and available on distro (Ubuntu or RHEL or SLES) (black), in community (purple), Lab7 (green) or IBM IOP (blue). Does not mean it is optimized. Does not mean that a commercial ISV version is available.

Evaluating: Needs to be vetted in new business development prioritization process. Some of these are available codes that need optimization to be competitive with x86.

Available:
Apache Web Server
Apache tomcat
Ceph, Chef Client, Glassfish, Jetty, Juju & Juju gui
Landscape client
MAAS, OpenStack, Puppet
Apache Qpid
Thrift
Ceilometer client, Sensu Server & Client

Port In Progress:
Chef Server v12

Available:
Alfresco, BTRFS, Bootstrap
CentOS
Chroma-key
Cluster Glue, corosync, CRMshell, DRBD
Evolution data svr, ClusterGlue, HAProxy, Heartbeat keepalived
Ldirectord, Linux-HA. mesos
OpenSSL
Pacemaker, REAR, resource agents, Samba, Tophat
WordPress

Port In Progress:

Evaluating:
Cluster-Network
Open Identity Stack (forgerock.com)

Available:
ALLPATH-LG, bfast, BioConductor, BLAST, BOOST, Bowtie2, BWA, EMBOSS, FASTA, FastQC, HMMER, HTSeq, ISAAC, MUSCLE, PICARD, PLINK, RNAStar, RSEM, SAMTools, SAMTools 1.0, SeqAn, SHRiMP, SOAPAligner, SOAPbuilder, SOAPDenovo, TMAP, TopHat, ABySS, Balsa, Caffe, Code-Saturne, GATK, GMP, Heat3d, IOR, Lattice-Boltzmann, LAPACK, OpenFOAM, waterman ....

Optimized & GPU enabled:
AMBER14, CHARMM, Galaxy, GAMESS, GROMACS, HPC-challenge, KKRnano, LatticeQCD, MAFIA, NAMD. NEST, PEASOUP, PLUTO, Quantum Espresso, QUDA, SOAP3-DP, WRF

Evaluating:
CEGMA, DIALIGN-TX, openQCD, SIFT, STREAM, miRdeep2, ucsctools, ViennaRNA,

Solution for Hadoop – Client/Industry Use Case

Academic and Business Partnership use case:
The Poole College of Management at North Carolina State University integrates teaching, research and active engagement with businesses to uncovering new business opportunities

Business challenge:
Find a solution that could efficiently handle an extremely large volume and variety of structured and unstructured data.
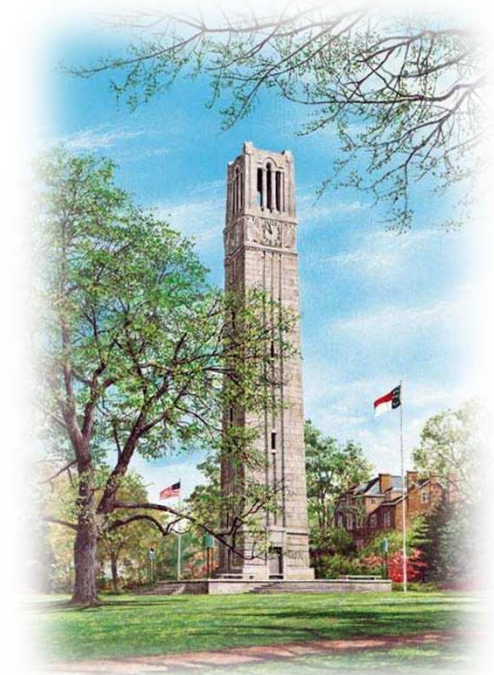
Solution:
Computing environment allowing students to run analytics models, including natural language text, on structured and unstructured data with IBM BigInsights and Power Systems.

37x FASTER INDEXING
14 days to 9 hours

3.5x LESS INFRASTRUCTURE
14 x86 servers to 4 Power Linux servers

Read NCSU Story: http://www.ibm.com/software/businesscasestudies/us/en/corp?synkey=T707869Y23501L49

NC STATE UNIVERSITY

POWER8

Customer Success Stories on POWER

A facilities management company in the UK improved their reporting..

7 hours to
2 seconds

Running our business warehouse on DB2 with Linux on IBM Power Systems has increased IT efficiency, reducing data backups . **99.9% improvement** reduced data volumes by up to **80 percent** performance optimization by **50 percent**.

15 hours to
5 seconds

Pharmaceutical Supplier :Negotiates multiple times per year with Medical Suppliers and Medical Aids. They have thousands of products, negotiations are very time constrained 1 to 2 hrs.

44min to run report per product. Now its 0.2 milliseconds. This has improved their ability to negotiate. Adding Millions to the bottom line

We took 10 days of sales figures across all of the 400 stores of a large pizza company and the customer was able to get a report in

10 days to
3 seconds

A nuclear risk assessment company in France converted

Hours to
10 seconds

With the business analytics accelerator we got the reports of a large men's clothing retailer in Denmark

51 minutes to
10 seconds

Are you waiting more and accomplishing less?

IBM Power Systems™ deliver the power demanded by the Waitless World.

Find out more about Power Systems >

IBM®