

Big Data: Delivering an Agile Infrastructure for Time-Critical Analytics in Telecommunications

Sponsored by IBM

Srini Chari, Ph.D., MBA
October, 2013

<mailto:chari@cabotpartners.com>

Executive Summary

Consumers are driving transformation throughout the telecommunications industry with their rapidly changing attitudes and behaviors including increased use of smart phones and social media. As a result, Telecom big data (CDRs – Call Detail Records and XDRs – Extended Data Records) are growing quickly in volume and variety. To compete effectively in the current industry climate, Communications Service Providers (CSPs) must develop a deeper understanding of their customer preferences to provide higher-value services to maintain customer loyalty and reduce churn.

This requires changes in a CSP's business processes and the increasing use of time-critical big data analytics. Yet, CSPs have limited budgets for major new infrastructure investments beyond what's already allocated to build 4G/LTE networks. For time-critical big data analytics, CSPs need a cost-effective and agile IT infrastructure to leverage their unique data assets to drive greater service innovation and deliver the best possible customer experience.

IBM recently collaborated with a major Asian CSP to benchmark one of their time-critical CDR query workloads associated with an external customer-initiated process of investigating consistency between a CDR and the associated charges. This workload included a data loading phase and query execution phase. One key requirement was to ensure a very quick response time even with hundreds of concurrent users. This CSP also had a limited budget so the entire solution including the infrastructure had to be very cost-effective.

*The IBM infrastructure solution, optimized for **Hadoop** and **MapReduce** workloads, consisted of IBM Platform Symphony and the IBM General Parallel File System (GPFS) running on IBM PowerLinux servers. This IBM solution performed much better than a functionally equivalent x86 infrastructure solution proposed by the competition. Compared to a competitor's x86-based appliance solution that only scaled up to 300 users before query response time became unacceptable, this IBM solution:*

- *Used only **half the number of servers**, helping reduce the total cost of ownership (TCO)*
- *Delivered **over 4 times the data loading performance per core** and*
- *Scaled well up to 900 concurrent users with **a query performance of 2-3 times better** across the range of CDR data set sizes analyzed.*

This IBM solution is the basis of the recently announced IBM Application Ready solution for InfoSphere BigInsights, a big data solution that provides an expertly designed, tightly integrated and performance optimized reference architecture for MapReduce-based big data workloads. This Application Ready Solution can easily be tailored to solve big data analytics problems in several industries including in Telecom.

Copyright © 2013, Cabot Partners Group, Inc. All rights reserved. Other companies' product names, trademarks, or service marks are used herein for identification only and belong to their respective owner. All images and supporting data were obtained from IBM or from public sources. The information and product recommendations made by the Cabot Partners Group are based upon public information and sources and may also include personal opinions both of the Cabot Partners Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. The Cabot Partners Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document. This document was developed with IBM funding. Although the document may utilize publicly available material from various vendors, including IBM, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury CT 06810, www.cabotpartners.com

Cabot
Partners
Optimizing Business Value

Competing with Real-Time Analytics in Telecom

CSPs typically have a very large subscriber base that is continuously generating lots of data. Data is generated with every call, text or transaction. Globally, data traffic is growing at a CAGR of 71%¹ to 10.8 exabytes per month largely because of the proliferation of smart mobile devices transacting in rich multimedia data such as videos, images and online games. CSPs must leverage this data to provide valuable actionable business insights. This often requires improved data management capabilities, new business models and associated charging mechanisms to outflank competitors.

For example, CDRs provide information about who subscribers called and their locations. The data also often provides detailed information about the customer experience and quality of service². Did the call go through easily? Was it dropped or interrupted? In addition, a new category of data, XDRs, are being captured from the rapidly growing use of smart mobile devices. XDRs typically provide information on which web sites subscribers visited, which apps were used and what transactions were executed. This data often provides detailed information about the response time for an app or song download or an eCommerce transaction.

CSPs also have a lot of internal data in the back office and customer interaction logs. All this voluminous data can be static or dynamic and exist in a variety of forms: structured (in existing enterprise IT systems like customer relationship management (CRM), sales force, and billing, etc.), unstructured (voice calls, videos, social media, email, and chats).

As carriers of data, many CSPs are improving their CRM systems by mining usage patterns to derive customer behavior, demographics, lifestyle and social influences and locations. But they still continue to face significant business challenges including threats from newer and more nimble entrants, significant erosion of their large voice based business to data based alternatives, stronger bargaining power of customers, and a demanding supplier base that continues to exert more pressure functioning as vendors rather than as strategic partners. This has resulted in escalating competitive rivalry and declining profits in the industry.

CSPs are beginning to differentiate their services² by being proactively customer-centric; continuously monitoring quality of service in real-time and offering agile and pre-emptive customer care. A few CSPs are also offering high-value services such as proactively intervening when calls are dropped to offer customers compensation or another reward to maintain their loyalty, or identifying subscribers caught in critical situations like automotive accidents so that local emergency services personnel and next of kin can be automatically notified, and so forth. This dynamic decision making ability will further streamline and enhance current CRM processes while delivering a superior customer experience.

¹ Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017,

http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

² IBM Institute of Business Value, “Analytics: Real-world use of big data in telecommunications: How innovative communications service providers are extracting value from uncertain data”, <http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03552usen/GBE03552USEN.PDF>

Call Detail Records (CDRs) and Extended Data Records (XDRs) are growing and can provide valuable insights

CSPs are competing by offering high-value real-time customer-centric services by analyzing CDRs

An Enterprise-Grade Agile Infrastructure is required.....

Traditional approaches of offline analysis or business intelligence with siloed data marts cannot keep pace with this volume, variety and velocity of data that operators have to deal with today as a result of next generation network roll-outs, proliferation of smart phones and the rise of mobile social media. Technical obstacles include slow data loading and querying, large network latencies and the complexities of managing distributed infrastructure.

In order to deal with these preceding challenges and at the same time deliver new, revenue generating customer-delighting services without overloading networks or letting costs go out of control, CSPs are deploying innovative solutions to run their processes with fine-grained service control. This can be achieved by the intelligent use of analytics, particularly big data analytics. Analytics not only provides insight about existing business behavior but can also deliver models that can predict future outcomes far more accurately.

Innovative CSPs have begun to implement a high-performance, scalable and agile information foundation to support near real-time analytics capabilities. These include the use of emerging open source technologies such as Hadoop and NoSQL to reduce the processing time for the growing volumes of data especially in distributed computing environments. But they also need robust Reliability-Availability-Serviceability (RAS), security and governance processes, normally found in enterprise-grade IT solutions provided by companies such as IBM. By sourcing these IT solutions from IBM, CSPs will get a trusted partner and support throughout their big data analytics implementation journey.

....IBM provides High-Performance Big Data Solutions

IBM offers a wide array of leading [technical computing and big data solutions](#) including high-performance systems, clusters, HPC clouds, and pre-integrated big data solutions. Featured systems include: [IBM Power Systems](#) and [IBM System Storage](#) on Linux and IBM AIX. Key software includes powerful and intuitive workload and resource management software from [IBM Platform Computing](#), a high-performance shared-disk clustered file system - [IBM GPFS](#) and optimized scientific and engineering libraries. Additional innovative offerings include [IBM InfoSphere BigInsights](#) – a comprehensive, enterprise-grade full-featured Hadoop platform for Big Data analytics; middleware and business partner applications and service providers with deep proven expertise in the telecommunications industry. IBM also has a worldwide technical staff of domain experts to collaborate with CSPs to migrate and optimize their applications on IBM systems and software to solve their largest and most challenging big data analytics problems.

The following case study leverages many of these solution components including IBM Platform Symphony, IBM GPFS and IBM PowerLinux servers and is the basis of the [IBM Application Ready Solution for InfoSphere BigInsights](#) and [reference architecture](#) for big data analytics.

CSP big data infrastructure challenges include managing and optimizing performance on distributed systems and reducing latencies

Scalable and agile platforms that include Hadoop and NoSQL must be enterprise-class

IBM Application Ready Solutions for InfoSphere BigInsights are enterprise-grade and optimized for performance

Case Study: Delivering Outstanding Performance for an Asian CSP

IBM recently worked closely with one of the largest CSPs in Asia with over hundreds of millions of subscribers. A team from IBM Research and other IBM divisions benchmarked and optimized the performance of this CSP's CDR system on an infrastructure consisting of IBM PowerLinux servers, Platform Symphony and GPFS. This team reported outstanding performance and scalability for several uses of this CDR system, outflanking x86 alternatives from a competitor.

Customer Background: As one of the largest telecommunications companies in Asia with over hundreds of millions of subscribers, this company's primary business is to provide fixed line and mobile phone services. The company also provides broadband internet access to tens of millions of users and is the internet carrier with the largest Wi-Fi network in the country of operation.

- **Key Business Challenges:** This CSP needed real-time analytics capabilities to gauge the performance of their network, monitor usage patterns, detect potential fraud, settle customer disputes expeditiously and analyze subscriber credit risk. The lack of these real-time capabilities was impacting decision making and delaying customer response. This was leading to significant customer churn and dissatisfaction.
- **Key Technical Requirements:** Traditional relational databases are inadequate to fulfill the business demands of efficiently querying the growing volume, velocity and variety of CDRs and XDRs. A big data solution based on Hadoop and NoSQL is required. This CSP wished to explore a big data high-performance analytics solution that could concurrently query CDR data extracted as a raw data file and process this data in parallel by loading the data through multiple nodes on a file system similar to the Hadoop Distributed File System (HDFS). This is becoming a ubiquitous test case in Telecom.
- **Key Performance Requirements:** CDR data sizes are increasing dramatically. For a medium-sized CSP or a branch of a larger CSP, the CDR size is several 10s of billions of records/per month or numerous Terabytes worth of data to store and analyze. This size is expected to rapidly increase even more in the near future with the growth of new mobile internet high-value services. High concurrency and quick response time is also required. Each branch or location must support 1000 concurrent records/sec with a query response time of under a few seconds. Moreover, the infrastructure must scale-out linearly and be cost-effective as the CDR sizes grow.

Proof of Concept (PoC) Description: This PoC included collection of data (a solution built with the help of a local IBM business partner and ISV), loading the raw data on the file system and running several concurrent real-time queries. Two architectures were explored:

- **IBM System Architecture:** Figure 1 depicts the IBM infrastructure used for this PoC consisting of three PowerLinux servers running Platform Symphony and GPFS for the data loading and query software infrastructure. Each node is an eight core Power7 processor that is connected to an external direct attached storage with three drawers, for

Major Asian CSP plagued with slow data loading and customer query response resulting in significant churn

CSP big data PoC setup to analyze query response for hundreds of concurrent users using Hadoop and NoSQL

Performance must scale linearly with increasing data size and quick query response time in few seconds needed

IBM solution:
Fewer
PowerLinux
servers +
Platform
Symphony +
GPFS

Competitor:
Functionally
equivalent
solution
needed more
x86 servers

Benchmark
examined data
loading and
query response
performance
over a range of
CDR data sets

a total of 24 cores for the three servers. Each of the storage drawers was connected through Gigabit Ethernet with a capacity of 14.4 Terabytes.

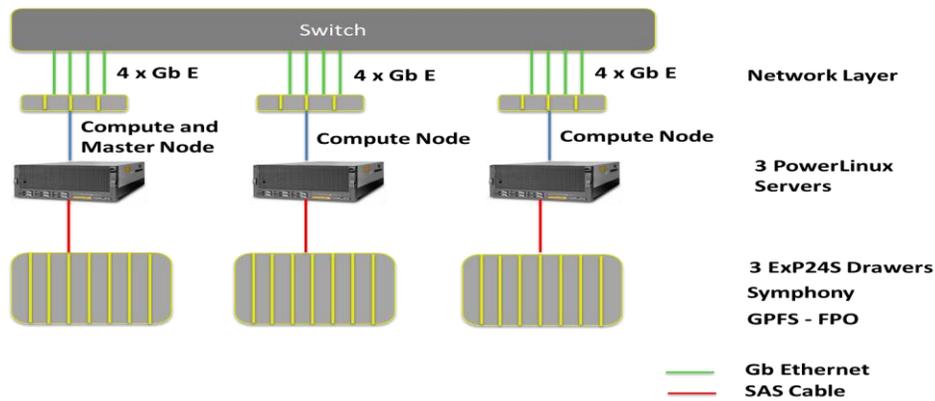


Figure 1: Proof of Concept IBM System Architecture

- Alternative x86-based Architecture (IBM Competitor):** Figure 2 depicts the competitor's alternative solution stack alongside the IBM stack. The competitor's solution consisted of six x86 servers each with 12 cores/server for a total of 72 cores. This solution used HDFS, HBase and substitute software in the competitor's appliance for the data loading and query infrastructure.

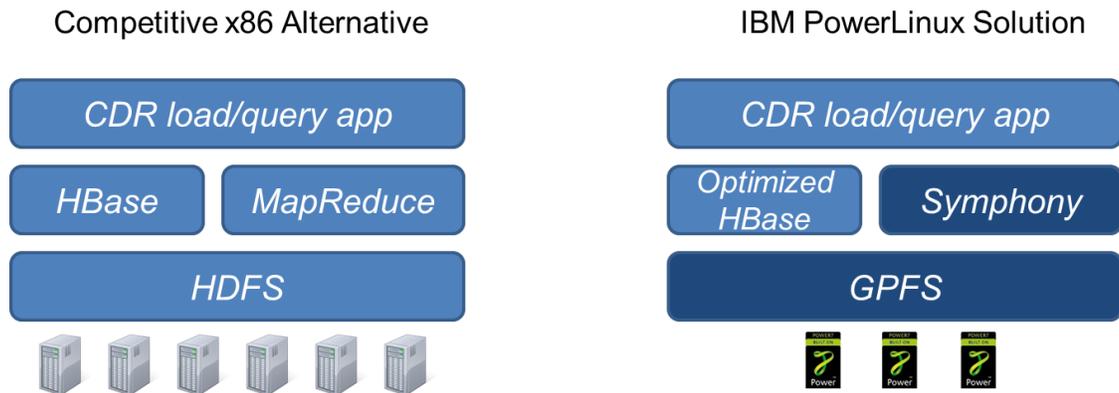


Figure 2: Alternative x86 and IBM PowerLinux Solution Stack

- Benchmark Description and Workflow:** For the IBM solution, the Hadoop cloud CDR load/query ISV application used an optimized version of HBase and IBM Symphony configured on top of the GPFS file system running on IBM PowerLinux servers. The competitor's solution ran the same CDR load/query scenario but with HBase and MapReduce on HDFS running on six x86 systems. The benchmark had two phases:

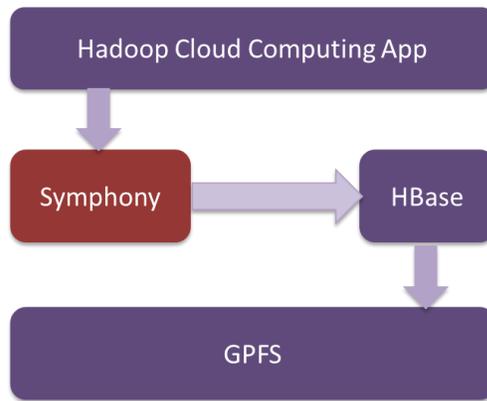
Phase 1 - Data Loading: The CDR data was exported from the client's billing system to the file system - GPFS or HDFS. In the IBM solution (Figure 3), Symphony was used to sort and load the data in parallel to the HBase repository which was stored on GPFS. In the competitor case, the data was sorted by either User ID or at random and then loaded to a HBase repository stored on HDFS. Figure 3 also shows the IBM and competitor results for data loading performance.

IBM solution was over 4 times the per core data loading performance compared with the competitor

CDR query execution response time scaled linearly on IBM PowerLinux servers which performed much better than the competitor

With more number of concurrent queries, the IBM performance, scale and TCO advantages with half as many servers become more evident

Data Loading - Flow



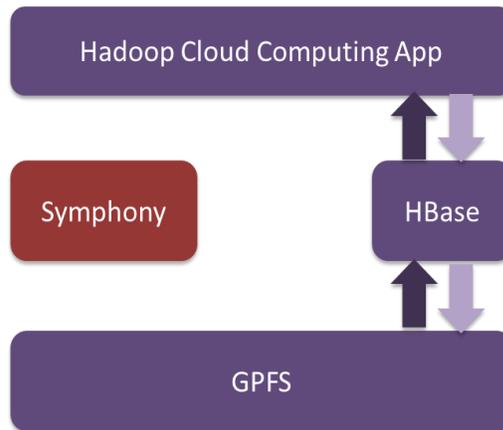
Data Loading Test Results IBM vs. Competitor

	IBM PowerLinux Solution with Platform Symphony and GPFS	Competitor X86 System
Hardware	Three PowerLinux Servers, self-attached SAS HDD, GigE network	Six x86, self-attached SAS HDD, GigE network
CPU Cores	24	72
CDR Loading Performance	300 K CDR/s	200 K CDR/s
Performance per Core	12.5 K CDR/s	2.7K CDR/s
Performance Ratio per Core	4.63	1

Figure 3: IBM Performs > 4X Better than Competitor for Data Loading Workflow

Phase 2 – Query Execution: Perform concurrent CDR queries on the data loaded in Phase 1. Each combination query used three parameters: user’s mobile number, start and finish time. Then the query result was displayed and the query response time was recorded. This was then repeated for a range of concurrent users. Figure 4 shows the workflow for the IBM solution that leverages GPFS alongside the query execution performance results.

CDR Query - Flow



Query Execution Performance IBM vs. Competitor

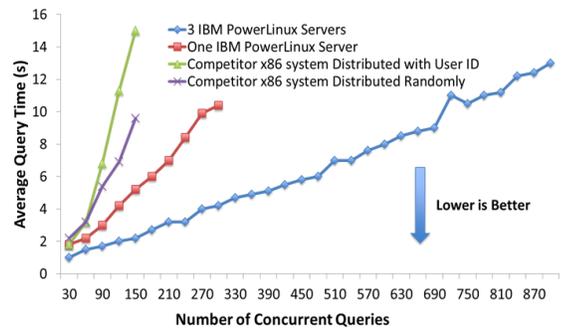


Figure 4: Query Execution Workflow - IBM Performs 2-3 X Better than Competitor

Proof of Concept (PoC) Results: The results clearly indicate the advantages of the IBM solution that significantly outperforms the competitor solution, but uses only half as many servers. Specific IBM performance advantages delivered over the competitor include:

- Over 4 times the data loading performance per core
- Good scaling well up to 900 concurrent users
- Query performance of 2-3 times better across the range of CDR data sets.

IBM Solution with 1/2 the Number of Servers is 4X Better in Data Loading and 3X Better in Query Performance

	Competitor x86 system	IBM PowerLinux Solution with Platform Symphony and GPFS
Database	HBase	Optimized HBase
Distributed Computing Framework	MapReduce	Symphony
File System	HDFS	GPFS
Operating System	Linux	Linux
No. of Servers (less is better)	 Six x86 Servers	 Three PowerLinux Servers
Data Loading Performance per Core		 4x
Query Performance		 3x
Scaling	 300 concurrent users	 900 concurrent users

Figure 5: IBM Solution with 1/2 the Number of Servers is 4X Better in Data loading and up to 3X Better in Query Performance

With this IBM solution, CSPs get a solution that not only delivers outstanding performance but should also lower the total cost of acquisition and ownership (TCA and TCO). Having fewer servers often translates to lower operating costs for facilities, electricity and labor.

Unique Business Value from the IBM Solution: We believe that this IBM solution is well-optimized for these time-critical CDR analytics workflows because the solution is:

- **Best of Breed.** Each solution component is optimized to handle its individual tasks very effectively with excellent performance and reliability.
- **Seamless.** Solution components complement each other, making the workflow seamless and fast.
- **Conformant with Common Open Industry Standards.** These IBM components support common open-source industry standards, making solution deployment simpler.
- **Supported End-to-End.** Customers get the benefit of end-to-end IBM support and services for the IBM components and overall solution

Benefits of IBM Solution Components vs. the Competitor’s Offerings are:

- **PowerLinux servers vs. x86 Linux servers:** Recent studies^{3,4} across a range of enterprises indicate that IBM Power systems perform better than x86 systems in RAS, performance, TCO, security and overall satisfaction. These results reinforce the fundamental advantages of the POWER7+ architecture, including the ability to support high levels of concurrent error detection, fault isolation, recovery, and availability.

³ Edison Group, “Better Performance, Lower Costs The Advantages of IBM PowerLinux 7R2 with PowerVM versus HP DL380p G8 with vSphere 5.1,” <http://public.dhe.ibm.com/common/ssi/ecm/en/po03161usen/POL03161USEN.PDF>

⁴ Solitaire Interglobal, “Why OS Matters”, <http://public.dhe.ibm.com/common/ssi/ecm/en/po03099usen/POL03099USEN.PDF>

IBM solution is optimized for a seamless high-performance workflow conforming to open standards and fully supported

Compared with competitor's offerings, IBM solution components provide higher performance, better RAS and efficiency and can support many concurrent applications at multiple locations

In addition, the Power7+ design delivers more balanced performance especially for data-intensive workloads such as analytics. Specifically, Power7+ has more threads per core (4 vs. 2), deliver higher throughput because of larger memory and superior I/O bandwidth. Further, it supports a more optimized implementation of the Java Virtual Machine (JVM). Advanced virtualization with PowerVM can be used to consolidate many different workloads/tasks including production databases, applications, high availability servers, backup/recovery servers and other servers onto a single, smaller set of servers. It is much easier to optimize parallel tasks on fewer servers than on the many servers that x86 alternatives will undoubtedly require.

- **GPFS vs. HDFS:** GPFS is an enterprise-grade parallel file system, fully POSIX compliant (so no need to pre-load data), delivering very high performance with no single point of failure. By contrast, to ensure high availability, HDFS requires an additional file-copy stage that further reduces performance.
- **Hadoop Accelerated by Platform Symphony vs. Native Hadoop:** Symphony is more reliable, efficient, and enterprise-grade with fail-over capabilities. It supports many concurrent applications and is ideal for more powerful server environments where consolidation can greatly increase server utilization and efficiency.
- **Optimized HBase:** IBM's deep technical client team provides unique capabilities to optimize HBase to achieve higher performance.
- **IBM BigInsights vs. Open-Source Hadoop:** providing full compatibility with Hadoop, BigInsights also provides unique capabilities and plethora of benefits that other frameworks do not offer, including:
 - Rich connectors to existing data sources
 - Tools and accelerators, to simplify and speed application development
 - JAQL - an easy-to-use, SQL like query language for manipulating both structured and unstructured data
 - Bigsheets, a spreadsheet-like tool that lets end-users easily analyze massive data sets without the need to write scripts or code.

Driving More Business Value for CSPs using Time-Critical Analytics

Innovative CSPs, like the Asian CSP mentioned earlier, are taking a pragmatic evolutionary approach by leveraging the valuable nuggets of information available in existing CDRs and other internal data. Using the IBM Application Ready Solution for InfoSphere BigInsights, CSPs can:

- Simplify and improve the performance and throughput of other CDR query applications
- Build a stable and sustainable analytics application development environment
- Reduce system development, implementation and maintenance costs
- Respond to many other real-time business agility needs.

Then, using another appropriate instance of the IBM Application Ready Solution for InfoSphere BigInsights, CSPs can mine XDRs captured from the rapidly growing use of smart mobile devices. Potential benefits from these additional insights include:

Innovative CSPs are first implementing analytics solutions using CDRs and then plan to leverage XDRs

More Effective Real-time Targeted Promotions: The attitudes and behavior patterns of customers are evolving so fast that most CSPs lack the agility to respond. Customer loyalty to their service providers has never been weaker, and the effectiveness of current customer promotions including those for Value Added Services (VAS) is being strongly challenged, especially in many growth markets where most subscribers use pre-paid mobile services with almost no switching costs. With real-time big data analytics solutions⁵, CSPs can deliver the right offer to the right customer at the right time, and in the right context, while offering optimal incentives to both targeted customers and dealers or partners. This can increase campaign effectiveness, improve revenue and customer experience, while reducing customer churn and potentially increasing customer loyalty.

Better Designed Products and Services: New products and services offered by CSPs are currently based on elaborate, non-operational market surveys that do not factor in the most current customer preferences or sentiments. Even the launch of these new products and services involve using traditional push marketing approaches that are less effective with today's demanding informed customers. CSPs are now collecting and analyzing enormous volumes and varieties of data on customer preferences and sentiments across multiple channels, venues and devices. This helps further enhance the design and launch of new products and services that are better tailored to the precise location, segment, demographics and geography of prospective subscribers. For example, a service with reduced roaming costs tailored for frequent travelers can simultaneously build loyalty and increase usage.

Deeper Network Analytics: With big data solutions to analyze CDRs, CSPs can improve their Network Analytics. This helps identify and resolve network bottlenecks in minutes, allowing CSPs to proactively manage customer experience and churn. CSPs can also better manage and plan their capacity requirements to improve the quality of service and optimize network investments to maximize benefits for high-value customer segments.

Location Based Services: The proliferation of smart phones presents new opportunities and challenges: consumers want the best deals for all purchases based on their current location while requiring the CSP to honor their privacy preferences and provide only relevant offers when requested or opted-in. Big data analytics solutions leveraging XDRs can combine lifestyle and mobile profiles with subscriber usage and digital behavior to create targeted high-value Location Based Services.

IBM: Paving the Way for Enterprise-Grade Solutions for Big Data

Just like the Asian example, many CSPs are already realizing significant business value by initially applying real-time analytics to mature, well-understood internal data and CDRs, and then to XDRs. These pioneering initiatives require capabilities for real-time actionable high performance analytics and an associated agile, enterprise-grade infrastructure for big data.

The IBM Application Ready Solution for InfoSphere BigInsights (Figure 5) is just that. It provides an expertly designed, tightly integrated and performance optimized architecture for MapReduce-based big data workloads and comes with:

- **Complete cluster:** A comprehensive, tightly integrated cluster designed for ease of procurement, deployment, and operation with all required components for big data applications; including servers, network, storage, operating system, management software, Hadoop compatible software and run time libraries.

⁵ <http://public.dhe.ibm.com/common/ssi/ecm/en/tls03054usen/TLS03054USEN.PDF>

*Benefits are:
highly effective
targeted
promotions,
better products
higher-value
services and
improved
customer
experience*

- **Workload optimized configuration:** The cluster is configured to maximize application performance and lower TCO. The cluster is optionally integrated with Platform Symphony and HPC, GPFS and InfoSphere BigInsights to improve time-to-value.
- **Flexibility to customize:** The hardware and software components in the cluster are customizable for the best performance or the best price/performance ratio.
- **Commercial, solution-level support:** The configurations are validated and supported for all hardware and software components.

IBM
Application-Ready Solution for big data includes a workload optimized cluster that can be customized to deliver the best performance or the best price/performance ratio for many industries

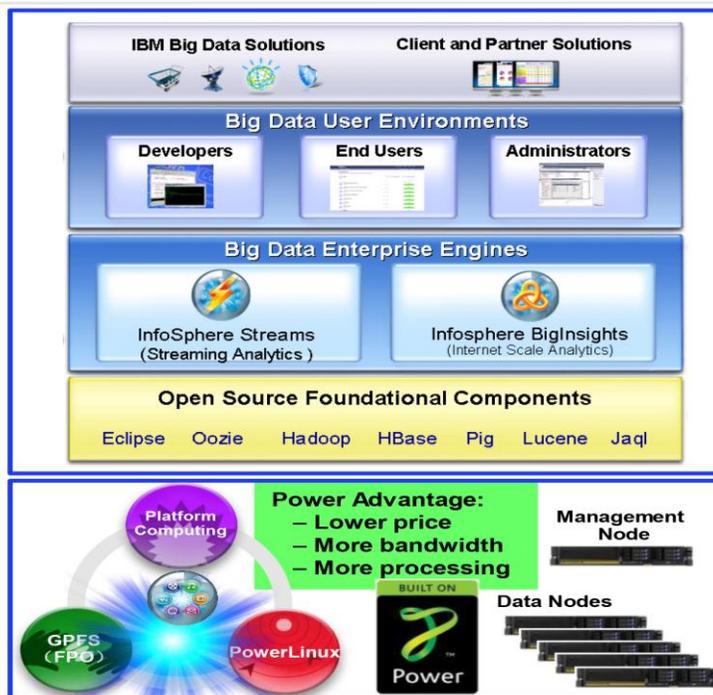


Figure 6: Solution Stack: IBM Application Ready Solution for InfoSphere BigInsights (source: IBM)

This IBM Application Ready Solution can also be used to solve similar time-critical data-intensive analytics problems in the financial, life sciences/healthcare, energy, earth sciences, manufacturing, retail and of course, telecom sectors.

For More Information

Please contact your IBM representative or visit the following:

IBM Technical Computing & Big Data: www.ibm.com/technicalcomputing/bigdata.html

IBM Platform Symphony: www.ibm.com/platformcomputing/products/symphony

IBM GPFS: www.ibm.com/platformcomputing/products/gpfs

IBM Platform Computing: www.ibm.com/platformcomputing

IBM PowerLinux Servers: www.ibm.com/systems/power/software/linux/powerlinux/

IBM InfoSphere BigInsights: www.ibm.com/software/data/infosphere/biginsights/

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies including HPC, Cloud Computing, and Analytics. To find out more, please go to www.cabotpartners.com.