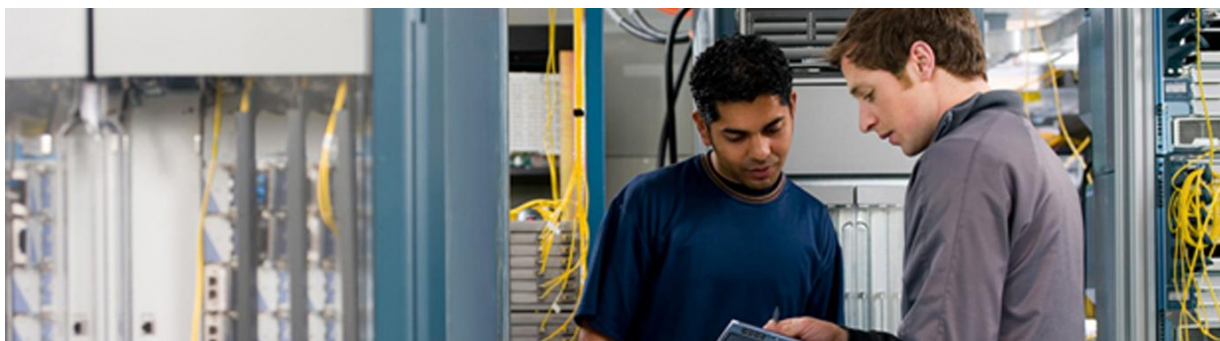


IT Workload Automation: Control Big Data Management Costs with Cisco Tidal Enterprise Scheduler



What You Will Learn

Big data environments are pushing the performance limits of business processing solutions. Platforms such as Hadoop can help analyze large amounts of complex, unstructured data, but require time-consuming management and can create service-level agreement (SLA) performance bottlenecks.

Cisco® Tidal Enterprise Scheduler provides workload automation that facilitates the flow of data between a wide variety of applications such as Hadoop, cutting costs and increasing the value realized from big data. Cisco Tidal Enterprise Scheduler provides:

- End-to-end business process management and visibility
- Capacity awareness of the entire business processing environment
- Real-time, latency-free job stream automation
- Reduction of errors in business analytics and reporting
- SLA awareness for historical and predictive analysis

Big Data Is Big Business

From data collection to storage, retrieval, analysis, and reporting, business processing depends on the ability to move data between files, applications, databases, and business intelligence solutions. For many years, advances in networking, storage, and data processing applications kept pace with the growing volume of enterprise data and the growing needs of business processing. But with the accelerated growth of unstructured data such as voice, text, and video streams, the traditional data processing approach breaks down. Analyzing the torrent of chat and text streams to measure customer sentiment is a different exercise than running a sales report against a Structured Query Language (SQL) database. Today's challenge is capturing unstructured data and quickly manipulating it to meet a variety of business goals and deliver valuable information to the end business user. This is the world of big data.

How Big Data Impacts Business Processing

Big data is the point at which standard SQL data storage and data warehouse processing breaks down. High data volumes and the speed of voice and video delivery are obvious causes. But there are other concerns as well.

- **Variability:** The type of data consumed may vary. Can it be stored in a standard SQL structure, or does it need a special application or dataset?
- **Variety:** Disparities may be due to the sources. Is the data coming in from different telemetry sources, and therefore formatted in different ways? Is it perhaps in Microsoft Excel format from one source and a data log or video stream from another source?
- **Scalability:** The ability to scale is a critical issue, as the growing volume and diversity of business data is the primary promoter of big data. Should the business processing system support linear scalability, or a different type of scaling environment? Can the infrastructure handle the quantity of incoming data? Are the CPUs sized correctly? Does the storage environment have the proper capacity and is it readily scalable?
- **Reliability:** Big data solutions must support 24-hour operation. Is the data processing software designed to withstand multiple failure points to support business continuity? Can the workloads be automated to such a degree that they do not create processing backlogs?

Gaming and continuously available entertainment data centers are good examples of big data environments that can put tremendous stress on data processing systems. Data flows into a casino data center from its website reservation system, security videos, and guest check-in with multiple data elements such as information about credit cards and unique customer needs. Some casinos also provide guests with a swipe card that records every activity, adding to the deluge of incoming data.

The data center must absorb this information 24 hours a day, 7 days a week, 365 days a year. There is no downtime to run batch windows; data must be processed and analyzed as it is collected. As an example, marketing professionals might want to analyze the profitability of existing offers, so they can provide guests with updated deals communicated through their in-room video.

Achieving this level of data management requires rapid analysis of huge streams of data gathered from a wide variety of sources. Without a smooth transition between the applications that manage the various types of input, bottlenecks and performance issues can reduce data center efficiency and prevent businesses from taking advantage of opportunities to increase revenues.

While gaming and always-on entertainment environments are examples of big data overload, big data is actually widespread.

- **Web and e-commerce:** Collecting, analyzing, and quickly using data is the primary focus for the Internet market. Web and e-commerce businesses need fast turnaround of collected data to understand customer behavior and turn it into delivered, targeted advertising.
- **Retail:** Like web and e-commerce businesses, retail operations also depend on fast data analysis of large amounts of structured and unstructured data to quickly turn data into actionable information. But point-of-sale transactions add the complexity of multiple touchpoints and multiple telemetry feeds, which complicate data storage, retrieval, and analysis.
- **Insurance and finance:** Protecting profitability while meeting customer needs is vital to these industries. Balance is critical. These businesses must make sure that they are creating the best possible offerings while limiting risk. Incoming data feeds must be collected and analyzed in near real-time to stay ahead of the market.

-
- Biotech and pharmaceutical: Regulatory compliance has a significant impact on business processing systems in these industries. Whether in manufacturing or in product trials, data streams from different sources must be timely and accurately recorded, processed, and analyzed. Every telemetry feed must be auditable. And as these companies move into multiple points of regulatory compliance, the tracking burden and the need for efficient big data management increase.

Regardless of industry, the goal of managing a big data environment remains how to capture vast amounts of diverse data, manipulate it, and quickly use it in a manner that is aligned with business requirements.

A New Era of Analytics

While structured data can be readily analyzed using traditional data mining techniques, unstructured data cannot. Correlating massive amounts of structured and unstructured data to develop useful information requires a big data analytics engine. A variety of platforms are available to meet the specialized needs of analyzing big data.

Big data solutions range from enterprise-level commercial offers to free, open-source solutions. Some handle only structured data, such as SQL databases and data warehouses, while others are compatible with unstructured data as well.

As an open-source tool with the ability to manage unstructured data, Hadoop is a popular platform for managing big data. Hadoop has several building blocks. These include top-level abstractions such as loading data into or extracting it out of the Hadoop environment, processing data into different datasets, and a file structure that allows data to be quickly retrieved and analyzed in real time.

While solutions such as Hadoop provide many advantages, one of the common pitfalls is a lack of integration with existing data warehouse technology. Hadoop, for example, is a standalone environment in which data is manually entered and reports are manually extracted. It was not designed to be an automated environment for the enterprise, and is not integrated with the broader business processing environments. As a result, there is no mechanism for monitoring and controlling the flow of data between enterprise resource planning (ERP) applications and Hadoop environments. Without proper workload management, any part of the infrastructure can be easily overloaded, reducing performance and reliability.

One Tool for Many Business Processing Touchpoints

With so much data coming in from so many sources and in such a diversity of formats, any business processing weaknesses can become magnified in a big data environment. As a result, a successful big data solution requires end-to-end business process visibility and a smooth transition between applications, databases, and big data processing environments. Big data workload solutions must therefore manage the input and output of data at every point through which data is acquired. Whether the data comes from telemetry feeds, reservation agents, retail consignments, or other sources, every piece of incoming data must be handled by one solution. And that toolset must have the following attributes.

- End-to-end visibility: A true enterprise workload automation solution that can scale to meet the demands of a big data environment must be able to see the entire business processing environment. Data sources – from business partners through FTP and data exchange mechanisms, to ERP and databases, to data warehouse solutions, to data integration, to business information solutions – need to be accessible through a single management console.

- Deep API integrations: Not only does the solution need a high degree of visibility, but it must be able to interact with each touchpoint at a comprehensive and easy-to-administer level. Each process step in the entire workload has to be smoothly integrated with the previous one, allowing administrators to quickly define, test, run, and customize the entire job stream. The tool must be able to interact with each application at a fundamental level without unnecessary code management or language barriers.
- Capacity awareness: The data center administrator needs to understand what percentage of data center resources each business process requires, and which tools the data passes through as it enters and exits each phase of business processing. Understanding the volume of data that will emerge from every input source is critical, because there is less and less opportunity to adjust to capacity fluctuations as the amount of data grows. A lack of capacity awareness can increase operating costs and processing errors, while decreasing resource utilization.
- Real-time, latency-free delivery: The ability to run business processes in real time is an important complement to capacity awareness, because latency is critical to the delivery of big data SLAs. As each process completes, another must pick up the data and continue the business process flow. This is traditionally a labor-intensive, manual process that must be automated in a big data environment.
- Real-time, error-free analytics: Analytics are the logical delivery end points of big data business processing environments. Lack of integration with business intelligence and analytics applications can result in a high error rate due to disconnected processes and missed or degraded service-level agreements.

Controlling Big Data Workflows

Cisco Tidal Enterprise Scheduler acts as an OS-like management layer that automates and tracks the flow of data between the infrastructure layer of computer, network, and storage devices, and the data movement layer of Hadoop, ERP, web services, and other data feeds.

Controlling data flow between the infrastructure and applications simplifies business processing management and integrates the Hadoop environment into the broad range of business processes in the enterprise. The functionality is similar to controlling water movement through dams, reservoirs, pumping stations, treatment, and pressure regulation points. Water is permitted to flow as needed and processed as needed without flooding any part of the system. When the tap is turned on you get the right amount, with the right quality, at the right time.

Cisco Tidal Enterprise Scheduler provides enterprise-class workload automation by handling the complete, end-to-end business process environment. Integration is managed through APIs with a number of different application sets and databases, including Oracle, Informatica, IBM, Microsoft, and SAP. Each interface has been carefully developed and tested to work transparently with each application. Combining these third-party data management partnerships with Cisco's workload automation solution has the added benefit of providing a low-risk path for transitioning a Hadoop environment from a test operation to an enterprise-ready solution.

Conclusion

As big data proliferates, solutions such as Hadoop must be integrated with existing data management environments. Cisco Tidal Enterprise Scheduler is a dynamic, intelligent, efficient solution for integrating big data processing and analytics into an existing data center infrastructure. It offers compatibility with leading applications, capacity awareness to prevent bottlenecks, automated process orchestration, and complete visibility across the business processing environment.

For More Information

To learn more about Cisco products and services, please visit <http://www.cisco.com/go/workloadautomation>.



Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)