WWW.WIPRO.COM

WIPRO
Applying Thought

PREDICTIVE INSIGHT ON BATCH ANALYTICS – A NEW APPROACH

Floya Muhury Ghosh

Table of contents

Abstract

While various industry players are investing in data management systems, several companies globally, face an increasing challenge of system failures and outages resulting in unhappy customers, brand damage and revenue loss. To illustrate, in March 2013, the regulatory body of U.K. fined a multinational banking and financial services holding company for failing to keep updated information on client objectives, risk profile and risk appetite. The regulatory body found that the reason attributed to this was the failure of the bank's data processing systems to allow sufficient client information to be processed and not so much a human failure. Such failures can have long term impact on companies with loss of potential customers even in the future. Out of several factors that can be attributed; failures and delays during data warehouse batch runs are an important area of focus for data managers today. This paper proposes a reference architecture to improve operational IT analytics with predictive capabilities to overcome such challenges. The proposed solution is vendor agnostic and provides consistent experience across a range of Data Integration and Business Intelligence tools.

Industry Landscape

Data warehouse and application service loss can adversely impact businesses in many ways like delay in financial closure leading to liability, penalty for noncompliance in delivering data on time and down time for information workers. On average, businesses lose between $84,000 and $108,000 (US) for every hour of IT system downtime, according to various estimates from the studies performed by industry analyst firms . Application problems are the single largest source of downtime, causing 30% of annual downtime hours and 32% of downtime cost, on an average . The leading cause of application downtime is software failure (36% of cost on average), followed by human error (22%) and third such rising costs are attributed to complex data warehouse environments that exist in companies today. The dynamic nature and complexity of the data warehouse environment can be attributed to

- Exponential growth of data
- Concurrent source and schema updates
- Decrease in batch load window and increase in real time processing
- Increase in data warehouse usage and demand for greater availability and performance

The increased complexity of the data warehouse environment intensifies the task of data managers manifolds. The ask of troubleshooting under such a dynamic environment becomes tougher. Additionally, a host of other issues such as human error, hardware failure, and natural disasters can disrupt data warehouse availability.

Over several years, innovative solution providers have tried to address these problems. There are multiple analytical solutions (Operations Management (OM) tools) existing today which help data managers monitor and measure the data warehouse environment. These solutions are catered to by features within Data Warehouse DI or BI tools and also within infra monitoring schedule management tools. The solutions help identify issues, isolate causes and resolve outages. The solutions also support performance management through IT infrastructure support.

However, these features which are restricted only to each particular tool, do not take into account the business process aspects and at best provide insights into what has gone wrong (instead of finding what would go wrong). These tools do not have the feature to constantly and proactively monitor system behaviour and provide real time insights into system performance and capacity trends. A typical solution today would analyse historical data to develop a certain trend in performance, after the failure has occurred. Whereas data managers are increasingly demanding capabilities to predict a failure before it occurs. This has created a need for predictive analytics driven proactive monitoring of data warehouse processes.

[1] http://www.bankingtech.com/182841/conduct-risk-explained-fix-the-systems-or-pay-the-fines/

[2] http://www.strategiccompanies.com/pdfs/Assessing%20the%20Financial%20Impact%20of%20Downtime.pdf

[3] http://www.prnewswire.com/news-releases/large-companies-lose-36-of-annual-revenue-to-network-downtime-infonetics-research-says-58973507.html

Current OM Tools – Limitations

As discussed above, a host of OM tools and solutions are used by data managers today. However, these solutions have the following limitations:

- Issue identification is reactive and not proactive – The current solutions largely work on a reactive approach by performing root cause analysis once the system failure has occurred. However, data managers require capabilities of the solutions to predict a system failure before it has occurred.
- Inconsistent view of business process – The data environment is ever growing and has multiple systems, devices, services, and applications. The current solutions do not provide a holistic view of business process failures which data managers require.
- Inconsistent experience across tools – The current solutions are limited in their capability to provide a consistent experience in integration, monitoring and data visualisation, thus limiting data managers' ability to garner a comprehensive view of the data warehouse environment.

Current OM Tools – Potential Improvements

The limitations of the existing solutions, outlined previously can be plugged in through "predictive insight on batch analytics".The various capabilities of such a solution are outlined below:

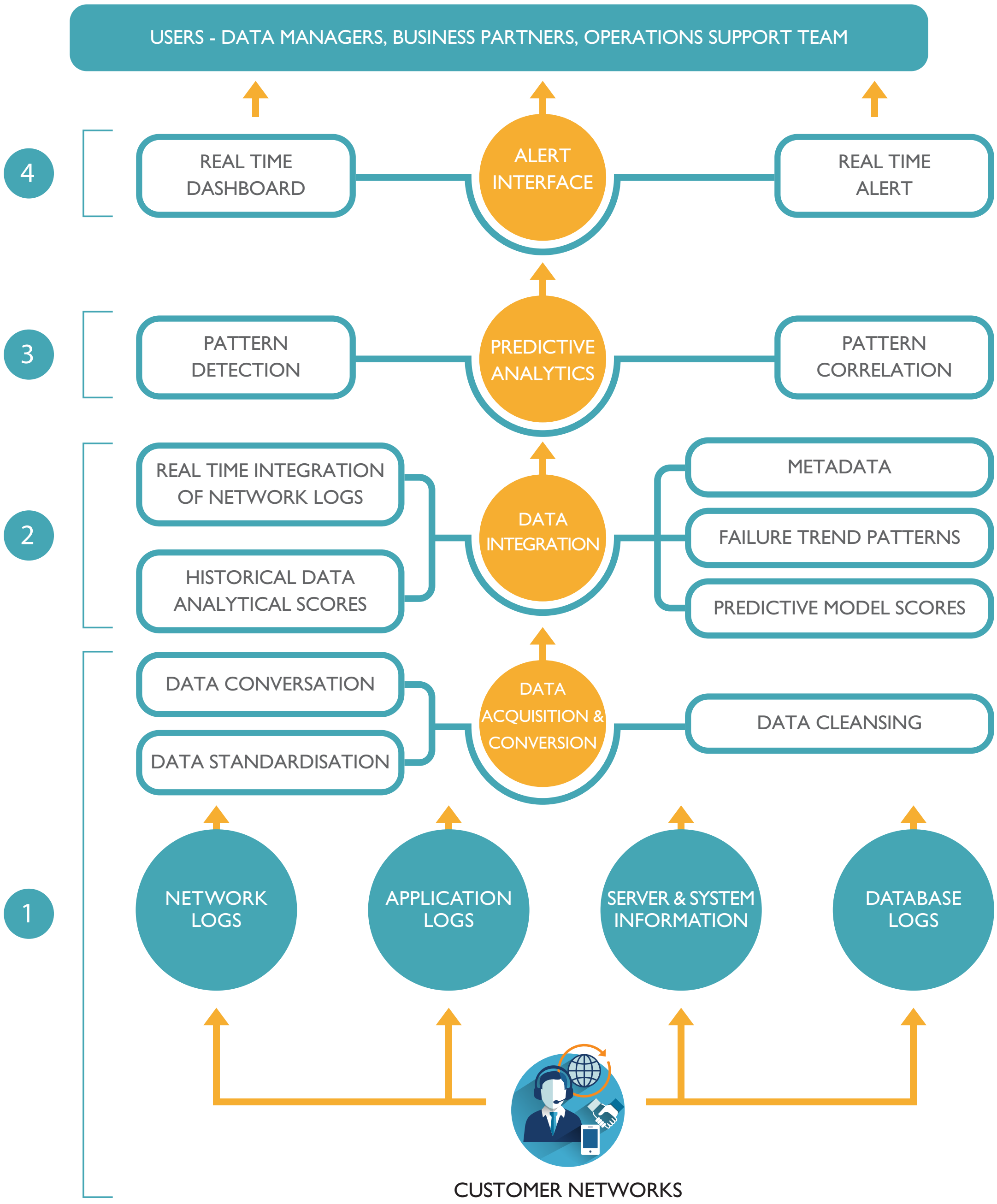
- **Predictive monitoring capabilities:** The solution should be able to proactively monitor the data warehouse environment based on historical as well as real time data processing.
- **Failure trend monitoring** – The solution should in specific monitor the failure trend and predict probable failures during data processing.
- **Troubleshooting in real time** - The solution should proactively troubleshoot issues from a business process data centre level view to a job / sub-task level view facilitating possible corrections before the system failure occurs.
- **Comprehensive view and monitoring of business processes:** The solution should provide insights on data warehouse execution environment parameters in real time to identify impact on business deliverables which further entails that the solution should have a holistic view of the impact on various business deliverables.
- **Integration of tools and technologies in the data warehouse environment:** The solution should effectively monitor and capture the environment

parameters across various data warehouse tools and technologies to provide a complete view of the data warehouse environment. In summary, the solution should be tool agnostic.

- **Consistent visualisation:** The solution should provide consistent user interface and outputs across different data integration, reporting and scheduling tools. Such a feature should enable data managers have a uniform view of the otherwise heterogeneous data warehouse environment.

Proposed Reference Architecture

Based on the characteristics outlined previously, below is the proposed reference architecture:



The numbered areas in the above reference architecture are described below:

1. Data Acquisition and Conversion:

- Real time system and environment parameters from network logs, server, application logs and database logs are collected from customer networks. The data is converted to a standardised format which makes the proposed architecture vendor agnostic and this enables seamless integration with other tools and technologies in the data warehouse environment while troubleshooting at real time.

2. Data Integration

- The network log data is integrated and correlated with knowledge based predictive models based on historical data. Historical and failure trend analysis is applied on patterns identified from network data.
- The proposed reference architecture also detects anomaly and failure patterns from network data.
- The proposed architecture correlates events identified with contextual insights like data volume, system availability, response time, processing time etc.

3. Predictive Analytics

- The architecture uses predictive analytics to quantify expected data processing behavior at real time. It uses multi-variate linear regression statistical models of the form $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$, $i = 1 \ldots\ldots n$ to achieve this. Some of the variables under consideration are job run duration, cache size, record count, throughput, CPU utilisation etc. The architecture also proposes to use Auto-regressive models to capture time variant processes and trends. The characteristic of the architecture thus, significantly differs from existing OM solutions primarily in its predictive capabilities which lie at the center of the proposed solution.

- A repository of business areas are also defined within the proposed architecture which enables the statistical model to predict the business area which would get impacted. This enables the data manager to easily identify the business area and drill down to the sub-task level to identify the potential impact.

4. Alert Interface

- The proposed reference architecture provides alerts and sets up communication channels on anomaly and next best action (NBA). The alert thresholds are configurable and are set based on system performance standards.

- Standardised inputs and uniform reporting interface provide consistent reporting across the data warehouse environment. Consistent reporting is a result of comprehensive view of the data warehouse environment that the proposed architecture enables.

Benefits

Several benefits can accrue to the data managers through the new approach:

- **Reduced Risks:** The new architectural approach de-risks batch operations through predictability, early warnings and fast issue resolution resulting in lower downtime of operations and improved reliability of data. In addition, infrastructure capacity utilisation forecasts and platform stability trends provide insight into long term risks.

- **Optimised Costs:** The proposed architecture further improves productivity by reducing manual monitoring and troubleshooting efforts resulting in reduced Total Cost of Ownership, lower cost of batch optimisation exercise and improved productivity of the platform due to reduced downtime and improved system availability.

- **Standardised Operations:** Data managers can avail standardised batch operation processes across technology stacks leading to – i) A tool agnostic approach-The alert management, predictive insights, diagnostics and trend reports are uniform irrespective of the tool used. This also helps reduce system upgrade costs. ii) Simplified operational processes and iii) Improved standardisation of IT operations overall.

Conclusion

This proposed architecture catering to improving operational IT analytics is vendor agnostic and provides consistent experience across a range of data integration and business intelligence tools. It provides predictive insights that are business process aware and helps achieve higher availability and better reliability of data environments. It combines ability to monitor infra resources, integration tools as well as job schedules. Data managers will find the proposed solution cost effective in addition to helping reduce overall operational risks and design standardised operations. It will help organisations maintain high customer centricity through improved operational processes.

References

1. http://www.bankingtech.com/182841/conduct-risk-explained-fix-the-systems-or-pay-the-fines/
2. http://www.strategiccompanies.com/pdfs/Assessing%20the%20Financial%20Impact%20of%20Downtime.pdf
3. http://www.prnewswire.com/news-releases/large-companies-lose-36-of-annual-revenue-to-network-downtime-infonetics-research-says-58973507.html

About the Author

Floya Muhury Ghosh has over 9 years of consulting and delivery experience in data integration tools and technologies. In her current role, she is a part of the Strategic Solutions team in the Data Integration practice at Wipro. Her present focus is in developing enterprise operational analytics solutions which would help IT operations to proactively act on batch jobs that could potentially fail thus helping in timely interventions thereby maintaining Business SLAs.

About Wipro Ltd.

Wipro Ltd. (NYSE:WIT) is a leading Information Technology, Consulting and Business Process Services company that delivers solutions to enable its clients do business better. Wipro delivers winning business outcomes through its deep industry experience and a 360 degree view of "Business through Technology" - helping clients create successful and adaptive businesses. A company recognised globally for its comprehensive portfolio of services, a practitioner's approach to delivering innovation, and an organisation wide commitment to sustainability, Wipro has a workforce of over 150,000, serving clients in 175+ cities across 6 continents.

For more information, please visit www.wipro.com

WIPRO
Applying Thought

DO BUSINESS BETTER

WWW.WIPRO.COM   CONSULTING | SYSTEM INTEGRATION | BUSINESS PROCESS SERVICES

WIPRO LIMITED, DODDAKANNELLI, SARJAPUR ROAD, BANGALORE - 560 035, INDIA TEL : +91 (80) 2844 0011, FAX : +91 (80) 2844 0256, email : info@wipro.com

North America Canada Brazil Mexico Argentina United Kingdom Germany France Switzerland Nordic Region Poland Austria Benelux Portugal Romania Africa Middle East India China Japan Philippines Singapore Malaysia South Korea Australia New Zealand