CTOlabs.com

White Paper:

Three Open Blueprints For Big Data Success

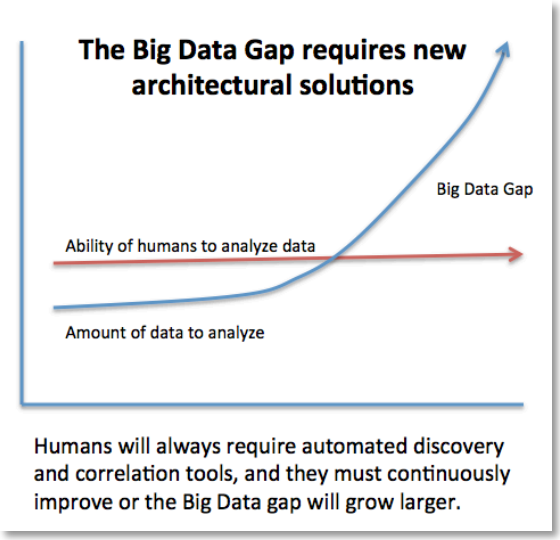
Featuring Pentaho's Open Data Integration Platform

Inside:

- **Leverage open framework and open source**
- **Kickstart your efforts with repeatable blueprints**
- **Tailor these use cases for your enterprise**

About This Paper

Enterprises are awash in more data than they can make sense of. This has given rise to the current “Big Data” phenomenon, where the opportunity for sensemaking over data calls for new solutions.



Federal enterprises and their leaders have been on the cutting edge of community efforts at big data, with most all agencies either executing on a comprehensive big data strategy or empowering technologist to explore and prove out possible solutions so a strategy can be developed.

Wherever your agency falls on the spectrum odds are very likely that you have established requirements for open framework and open repeatable solutions for your big data projects.

This paper is designed to facilitate a more rapid implementation of Big Data solutions by sharing three blueprints provided to the community by Pentaho.

About Pentaho

Pentaho delivers a business analytics framework based on open concepts and open source software, and they support open exchange of lessons learned by a vibrant community of users. Their capturing of successful use cases seen across industry segments led to their production of the Big Data Blueprints presented here.

About The Three Use Cases

The use cases presented here have been widely used in enterprises seeking to optimize their use of data to drive decisions. They are:

- The Case of Optimizing The Data Warehouse
- The Case of Streamlining Data Optimization/Refinement
- The Case of 360-Degree Customer Views

Optimizing The Data Warehouse

The data warehouse optimization (or sometimes referred to as data warehouse offloading or DWO) is a great starter use case for gaining experience and expertise with big data, while reducing costs and improving the analytic opportunities for end users. The idea is to increase

the amount of data being stored, but not by shoving it into the warehouse, but by adding Hadoop to house the additional data. Once you have Hadoop in the mix, the open framework of Pentaho makes it easy to move data into Hadoop from external sources, move data bi-directionally between the warehouse and Hadoop, as well as makes it easy to process data in Hadoop. Again, this is a great place to start. It's not as transformative to your business as the other use cases can be, but it will build expertise and save you money.

Pentaho simplifies offloading to Hadoop and speeds development and deployment time by as much as 15x versus hand-coding approaches. Complete visual integration tools eliminate the need for hand coding in SQL or java-based MapReduce jobs. The objective is to Save data costs and boost analytics performance.

Objective design features include:

- An intuitive graphical, no-coding big data integration.
- Access to every data source – from operational to relational to NoSQL technologies.
- Support for every major Hadoop distribution with a future-proof adaptive big data layer.
- Ability to achieve higher processing performance with Pentaho MapReduce when running in cluster.
- 100% Java, fast and efficient.

Here is an example of how this can look inside an IT landscape:

- An organization can leverage data from disparate sources including CRM and ERP systems.
- A Hadoop cluster will be implemented to offload less frequently used data from the existing data warehouse.
- Organizations will save on storage costs and speed up query performance and access to their analytic data marts.

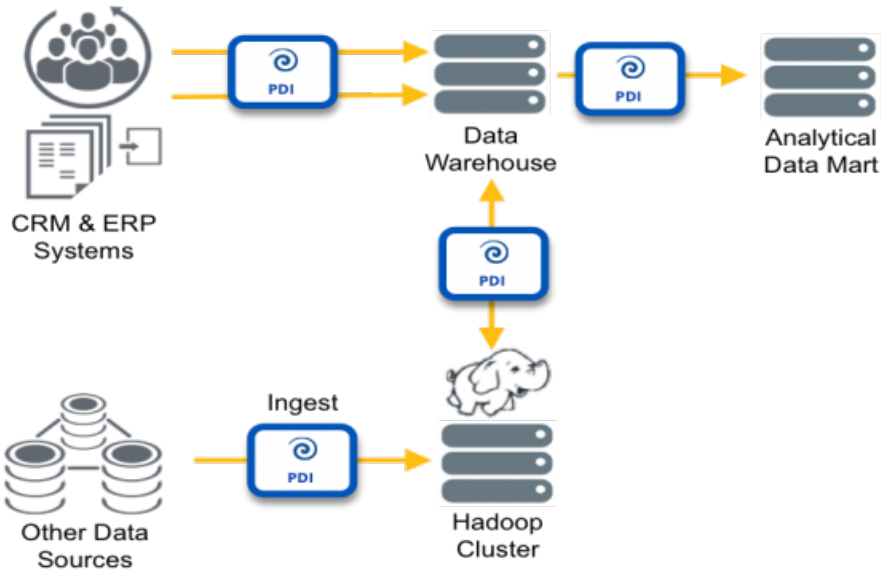


Figure 1: Optimizing The Data Warehouse

Return on investment of this approach can be measured by factors like:

- Staff savings and productivity: Pentaho's Visual MapReduce GUI and big data integration means existing data warehouse developers can move data between the data warehouse and Hadoop without coding.
- Time to value: MapReduce development time is reduced by up to 15x versus hand-coding based on comparisons.
- Faster job execution: Pentaho MapReduce runs faster in cluster versus code generating scripting tools.

There is no quicker or more cost-effective way to immediately get value from data through integrated reporting, dashboards, data discovery and predictive analytics. You should expect up to 15x data cost improvement with this approach.

Streamlined Data Refinery

The idea behind the refinery is to provide a way to stream transaction, customer, machine, and other data from their sources through a scalable big data processing hub, where Hadoop is then used to process transformations, store data, and process analytics that can then be sent to an analytic model for reporting and analysis. In many enterprises this is a logical follow on activity after optimizing data warehouses.

This can help you turn Hadoop into a Valuable Multi-source Business Information Hub, Just waiting to be queried. Pentaho's agile data integration and analytics platform allows you to stream data through Hadoop for transformation processing and immediately push the refined data to any analytic databases. For the end-user, a rich set of data discovery, reports, dashboards and visualizations are immediately available.

Objective design features include:

- Flexible data integration allows data to be seen as it is transformed shaving days/weeks off the development cycle.
- 15X faster than hard coding, Pentaho's GUI for MapReduce integration allows data to be moved and processed between Hadoop and ANY data source or system.
- Broad data integration accommodates and grows with your existing architecture.
- Powerful array of self-service analytics and visualization for all end users - business users, analysts, and data scientists.

A common use case can be best seen in the following example:

- An electronic marketing firm created a refinery architecture for delivering personalized offers.
- Online campaign, enrollment, and transaction data is ingested via Hadoop, processed and then sent on to an analytic database.
- A business analytics front-end includes reporting and ad hoc analysis for business users.

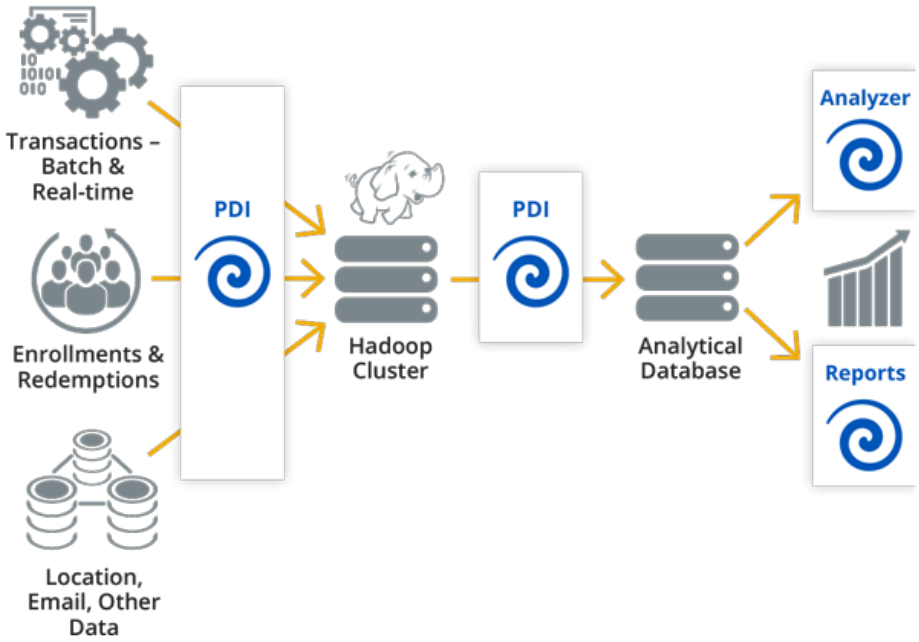


Figure 2: Streamlining The Data Refinery

Return on investment of this approach can be measured by factors like:

- Business users have immediate insight into ALL data
- IT can scale ETL and data management ensuring cost savings
- Engineer new data sets on-the-fly for prediction and trends
- Data driven insight into patron preferences are provided
- 80% reduction in processing time for faster insights

Customer 360-Degree View

For many organizations this can be a very transformative use case. For federal enterprises, analogies can be found for internal agency workflows and for agency-to-agency support. The idea here is to gain greater insight into what your customer or user of your information is doing, seeing, feeling and of course understand what they will need. All with the idea that you can then serve that customer better, therefore serving your organizational mission and the nation better. This blueprint lays out the architecture needed to start understanding your customer better. It will require significant effort in accessing all the appropriate customer touch points, but the payoff

can be huge. Don't worry too much about getting the full 360-degree view at first; starting with even one small slice can drive huge positive changes.

Integrating diverse data sources is simplified with Pentaho's broad support for both big and traditional data sources allowing the 360-degree view to be extended to external and internal customer related data. The Pentaho platform scales as business grows, enabling routing of governed blended, time-sensitive streams of data to be distributed to customer-facing teams – in real-time empowering more productive and profitable decisions.

Objective design features include:

- All customer touch point data in a single repository for fast queries.
  All key metrics in a single location for business users.
- Rapid time to value through drag/drop visual development for big data integration.
- Adaptive Big Data layer insulates organizations from evolving big data technologies.
- Intuitive and customizable dashboards.
- Sophisticated ad hoc slicing/dicing and rich data visualization.
- Distributed reporting capabilities for sharing information across teams.
- Data mining and predictive analytics tools for data scientists.
- Easily embeddable into operational software and applications.

Here is an example of how this may look within an IT landscape:

- An organization can ingest data from various sources into a single big data store, which is frequently the Apache Hadoop framework and/or MongoDB
- Data is processed and summarized at the customer unique ID level to build the 360-degree view.
- Accurate and governed customer data is routed to the appropriate analytics views for each role, including call center staff, research analysts, & data scientists.
- Using Pentaho, any data source is easily blended with an easy-to-use visual development environment for fast, simplified and stream lined integration.

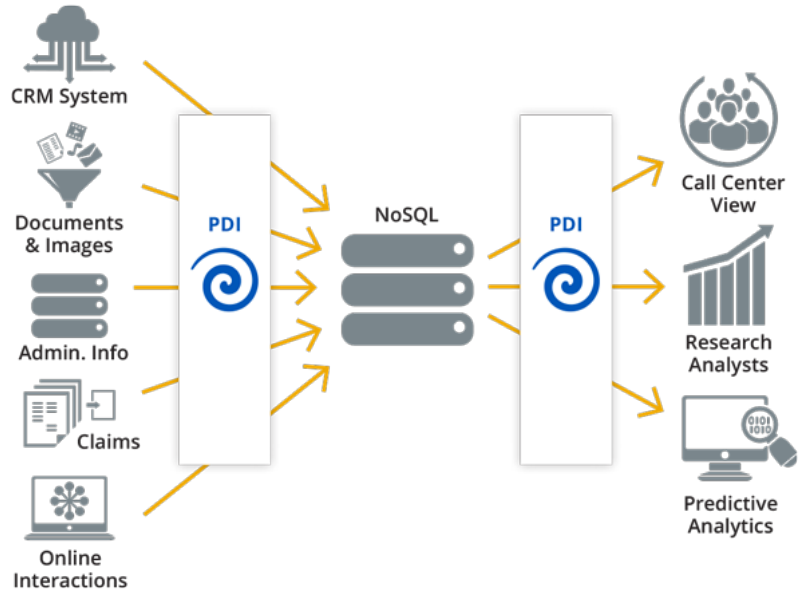


Figure 3: Customer 360-Degree Assessments

Return on investment of this approach can be measured by factors like:

- Staff Savings and Productivity: Rapid time to value through drag and drop visual development for big data integration.
- Operational Intelligence: Embed analytics into actionable line-of-business applications for each relevant customer-facing role.
- Reduced Risk: Protect data flow processes from changes in big data technologies with the Adaptive Big Data Layer.
- Instant Access to the Right Information for all Roles: Comprehensive analytics deliver easy to use ad hoc analysis, data discovery, advanced visualizations, highly formatted reports, and powerful dashboards.
- Reduced ETL time to analyze blended data from Hadoop, Hbase and data warehouse

Concluding Thoughts

The open Big Data Blueprints presented here can help you accelerate your projects by giving you repeatable frameworks you can tailor to meet your needs. We hope they help accelerate your implementation of enhanced data analysis capabilities and believe they can accelerate the use of data in support of your mission.

Please give us your thoughts on these approaches, we would love to have your feedback.

More Reading

For more federal technology and policy issues visit:

• CTOvision.com- A blog for enterprise technologists with a special focus on Big Data.

• CTOlabs.com - A reference for research and reporting on all IT issues.

• J.mp/ctonews - Sign up for the government technology newsletters including the Government Big Data Weekly.

About the Author

Bob Gourley is the co-founder of Cognitio and editor and chief of CTOvision.com He is a former federal CTO. His career included service in operational intelligence centers around the globe where his focus was operational all source intelligence analysis. He was the first director of intelligence at DoD's Joint Task Force for Computer Network Defense, served as director of technology for a division of Northrop Grumman and spent three years as the CTO of the Defense Intelligence Agency. Bob serves on numerous government and industry advisory boards.

For More Information

If you have questions or would like to discuss this report, please contact me. As an advocate for better IT use in enterprises I am committed to keeping this dialogue up open on technologies, processes and best practices that will keep us all continually improving our capabilities and ability to support organizational missions.

Contact:

Bob Gourley

bob.gourley@cognitiocorp.com

CTOlabs.com