WhereScape®

THE CASE FOR BIG DATA AUTOMATION

White Paper

WhereScape®

The Rise of the Logical Data Warehouse **01**
• Big Data Automation 02
What is Big Data Automation? **03**
• Ingenuity for Customer-Perceived Value 03
• Integrated BDA Platform 03
• Automated Data Integration 04
Making the Transition to Big Data Automation **05**
• Choosing your First BDA Projects 05
• Breeding Success 06
• Increasing BDA's Penetration within IT, Over Time 06
Replacing Artistry with Automation **07**

WhereScape®

THE CASE FOR BIG DATA AUTOMATION

The Rise of the Logical Data Warehouse

"The Logical Data Warehouse (LDW) is a new data management architecture for analytics which combines the strengths of traditional repository warehouses with alternative data management and access strategy. The LDW will form a new best practices by the end of 2015."

Gartner Group, Hype Cycle for Information Infrastructure (2012)

In 2011, McKinsey & Company announced the beginning of the Big Data Revolution, and suggested that the age of relational database management systems and SQL-based data manipulation and access methods was drawing to a close because those technologies could not keep pace with what McKinsey projected was a coming deluge of new and complex data sets, for most organizations in most industries. IT infrastructure suppliers — marketing, for the most part, open source technologies in the Hadoop ecosystem – rallied around McKinsey's projection, and tried – for a year or two – to unseat the RDBMS as the central component in enterprise-wide decision support architectures.

The Big Data industry's attempt to trigger a wholesale replacement of RDBMSs and SQL-based models with Hadoop, NoSQL databases and other, less useful, technologies, was not a success. But, when the smoke cleared in the industry at the end of 2012, a new architectural model had taken hold in organizations that needed to deal both with traditional, tuple-oriented data from transactional systems as well as with so-called "unstructured" (really, differently-structured) data sets entering the organization from outside: particularly market data, social media data, and large volumes of sensor-based data from what we are now calling the Internet of Things (IoT).

Gartner calls this new architectural template the logical data warehouse, and other analysts have similar models, with different names: the adaptive data ecosystem, hybrid enterprise data warehousing, and so forth. All of these architectural models and templates share an important – and correct – common impetus: for the foreseeable future, enterprise-wide decision support environments will be a highly-customized mix of traditional data warehouses and data marts, built-for-purpose data warehousing appliances, and Big Data technologies. We are, it seems, going back to the future, and returning to the best-of-breed, integrate-it-yourself practices that characterized our industry in the late 1990s.

Conceptually, logical data warehousing is a pragmatic, and a correct, approach to designing, building and operating large-scale commercial decision support systems. Big Data technologies – which, as practitioners are aware, include no mechanisms for data governance and precious little in the way of data security controls – cannot replace the reliability, security and manageability of the conventional RDBMS for any data sets that

WhereScape®

have governance, regulatory or compliance (GRC) facets. The multi-million seat installed base of SQL-based query, reporting and dash boarding tools cannot be retrofitted or adapted, easily, to work with Hadoop-based data pools. However, Big Data technologies do offer organizations attractive low-cost alternatives to traditional data staging and ETL processing environments, and also offer capabilities – for processing streaming data, for performing complex statistical analysis, and for machine learning, for example – that are not readily available in traditional RDBMS-based environments.

Combining the two ecosystems, and including proprietary data warehousing appliances like those from Oracle and Teradata where those for-purpose systems make business sense, is in the end the only credible architectural blueprint for the next ten years of commercial decision support systems design and implementation.

Contemplating the adoption of Big Data technologies, though, poses a problem for organizations that have recognized the wisdom of data warehouse automation (DWA). Big Data technologies are, almost without exception, completely artisanal: technologies made for, and by, people who hand-code, hand-configure, manually-inspect absolutely everything. No governance mechanisms. No metadata management. No configuration and change control. No security infrastructure. Those facilities – facilities we depend on, in modern decision support systems design and development – are assembled by hand, or done without. Big Data is a programmer's paradise; but that is the last thing most commercial IT organizations are looking for.

To make use of so-called Big Data technologies, it seems, we are condemned to embrace the oldest, least productive, and most failure-prone methods of building decision support infrastructure we know: hand-coding. Methods we have already, to our cost, learned can bury the IT organization in non-value-added maintenance, prevent it from responding in timely ways to ever-present end-user demand for the new.

So what is to be done?

Should we augment the automatable data warehousing infrastructure we've spend a decade or more building and tuning, with a new technology infrastructure that is radically different, and more expensive to build and operate than the environment we're in the process of automating now? Should IT organizations be automating their data warehousing environments, in order to redeploy expensive personnel to artisanal tasks in their big data projects?

Or should we learn from our mistakes, during the era of data warehousing, and build automation into our big data environments from the outset?

WhereScape believes that it's essential for organizations to pursue big data automation and data warehouse automation simultaneously: to automation the design, development, deployment and renovation of the entire logical data warehouse, in order to do much more, with much less, during a period of architectural transition and rapidly rising user demand that promises to last for a decade, at least.

Big Data Automation

Big Data Automation (BDA) is:

- a management discipline focused on ingenuity: applying IT talent and innovation to create value that is perceived, and rewarded, by internal customers
- a new way of thinking about how big data infrastructure and applications are designed, built and deployed in commercial environments
- an integrated platform of tools designed to automation routine IT tasks associated with designing, building, operating and modifying big data infrastructure and applications, or accelerating those tasks that cannot be completely automated

WhereScape®

What is Big Data Automation?

"Most of us are aware of business' unending frustration with IT's roadblocks to data.But, contrary to a current myth, the answer is not free for all, self-service access by users to every piece of data through some sexy visualization tool. That comes after the data is consolidated, blended, cleansed and certified for use. The creation of a high quality data resource has always been what data warehousing has been about. And that's what automation is about too—but faster, better and more flexible than traditional tools. With automation , we can move from IT's old need—or necessity—to control everything to empowering both business and IT people to each do what they do best. Business defines what data is needed and how it should be analyzed iteratively, with IT capturing the business needs and applying quality and production values in flight."

Barry Devlin

- part of a larger emerging movement within the technology industry, focused on automated data integration (ADI) across both traditional data warehousing and emerging big data infrastructure.

Ingenuity for Customer-Perceived Value

The essential feature of big data automation (BDA) is the realization that doing a lot more with a lot less is the new reality for IT organizations, and that smart IT organizations, in response to that reality, will seek to spend their scarce human talent — their collective ingenuity — building value that their internal customers within the organization perceive to be valuable: new, enabling, differentiating. Today, that ingenuity is trapped, because IT teams are spending most or all of their time performing routine internally-focused tasks, many of them invisible to internal business constituents (and therefore not perceived as valuable), and most of them either susceptible to automation (so that no human labor is spent performing them) or acceleration (so that less human labor is spent performing them). The first rule of big data automation is: spend human talent where it adds customer-perceived value.

Integrated BDA Platform

An integrated big data automation platform will include, at minimum:

- mechanisms for managing all of the metadata associated with the big data "data pool" — where data sets reside, how they entered the pool, what data engineering flows they are implicated in, what kinds of algorithms and decision-making processes the data sets are suitable (and unsuitable) for, what kinds of governance, regulatory and compliance (GRC) restrictions are associated with the data sets, and so forth. This is a technology domain entirely missing from generic open source based Big Data distributions, today.

- mechanisms for expressing, managing, versioning, executing and monitoring data

WhereScape®

engineering work streams — the kinds of tasks traditional hand-coded for MapReduce-based execution today

- mechanisms for expressing, managing, versioning, executing and monitoring the deployment of models and algorithms consuming data from the data pool and from data warehouses and marts, via industry standards like Predictive Modeling Markup Language (PMML)

- mechanisms for orchestrating data flows that cross the logical boundary between the data pool, and the data warehouse, regardless of where those flows originate, or terminate, or how often they cross that logical boundary.

- a studied neutrality with regard to underlying big data technology choices, and equal facility with batch-oriented big data infrastructure like Hadoop, and real-time stream-oriented infrastructure like Spark.

Automated Data Integration

Finally, big data automation is, strategically, an acceptance of:

- the critical important role that big data and its associated technologies will play in many company's future decision support infrastructure

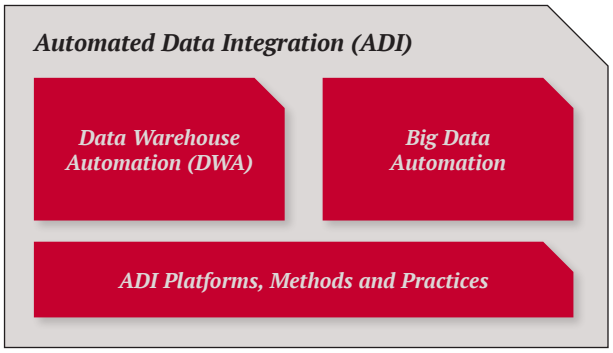
- a frank recognition that, in themselves, big data technologies require even more — not to mention more complex — manual work: to integrate the technologies, to configure them properly, to write the code that makes big data technologies actually work

- an understanding that the organization will be compelled — by market pressures, in most cases — to adopt some form of logical data warehousing architecture, for the foreseeable future, and be compelled to operate a complex, hybrid infrastructure in which conventional data warehouses and data marts, for-purpose data appliances and a growing data pool must cooperate to produce a low-cost, rapid-turn, high-efficiency enterprise decision-making environment.

As organizations shift their attention to Big Data and its associated technologies, responsive IT organizations will need to shift resources — requiring an even more rapid and thorough automation of conventional data warehouse tasks that would otherwise be the case. To participate in the organization's transition to Big Data and advanced analytics, IT organizations need to free up their best, most ingenious data-driven decision-making experts to work with the business units who are drawn to Big Data and its associated technologies, and who will be implementing those technologies and making use of that data, with or without IT involvement.

IT Automation pundits and practitioners also argue — correctly — that the application of automation disciplines to Big Data technologies is vital to the effective leverage of those technologies and their associated data sources: that there is a need, now, for "big data automation" or "data pool automation" or "data lake automation," just as there is for data warehouse automation. Big Data technologies, and Big Data itself, are in fact less well-organized, less well-integrated, less well-provisioned with metadata and less orchestration-ready, than conventional data warehousing technologies and conventional data sources. Big data technologies require, today, much more hand-coding than conventional data warehousing, much more documentation, and much more day-to-day intervention.

WhereScape®



The relatively clean, relatively well-documented, relatively stable environments we have enjoyed in conventional data warehousing are largely absent in the Big Data world; one cannot, when all is said and done, determine where particular data in the Hadoop "data pool" has come from, let alone whether that data is accurate, or safe for particular analytical uses. We cannot determine, without significant effort, what processes and groups are using a given data set for what purposes, or what impacts changes to, or loss of, that data set would have on the company's decision-making abilities.

These practitioners are beginning to talk about an automated data integration (ADI) market that brings the mindset and process orientation present in DWA to a larger environment in which data warehouses and data marts coexist with, and are integrated into, Big Data technology infrastructure. They point out — again, correctly — that metadata is metadata, regardless of the technology employed to persist data, that decision-making models are decision-making models whether they are embodied in dimensional schema or statistical analysis engines, and the problems that plague data warehousing today — too few people, too many demands, and not enough time or money — are almost certainly going to accrue in the Big Data world, as the technologies become commonly-implemented and therefore incapable, in their own right, of conferring competitive advantage.

Making the Transition to Big Data Automation

The keys to making a successful transition from artisan-style big data pilot projects to production-grade big data automation (DBA) include:

- choosing your initial project targets carefully
- breeding successes, within IT and within the business, through conscious internal communication, interaction and collaboration
- increasing the breadth and depth of the application of BDA methods and tools, within the IT organization, consistently, over time.

Choosing Your First BDA Projects

The key to building an effective — and cost-effective — logical data warehousing environment, in which mainstream relational infrastructure and Big Data infrastructure cooperate effectively, is to focus, early, on a big data automation project in which big data technologies serve as a staging area for data that is intended, eventually and in a perhaps-significantly modified form, to end up in the enterprise data warehouse or a particular data mart.
In other words, a good first BDA target project is one that exercises automation tools in a work stream that crosses the line between big data and enterprise data warehousing, and treats big data-resident data sets as sources for use with the established enterprise decision-making environment.

By contrast, using BDA tools to automate the sandbox or prototyping work of data scientists — who often work in a purely exploratory mode, with a very low yield of models and algorithms for productization and broad-based use — is counter-intuitive, and counter-productive, as sandbox

WhereScape®

and prototyping work is precisely the area in which artisanal methods and practices are optimal.

We find three common patterns, in logical data warehouses, that make for good initial BDA automation projects:

- Data sets resident in the data pool – native or re-engineered using Big Data programming constructs -- contain information that is of benefit to users operating in the conventional enterprise data warehousing environment, and must be included in the data integration work stream that produces populated schema in a data warehouse or data mart.

- Data sets resident in the enterprise data warehouse or a data mart – often seen as standard reference data, from within the data pool – need to be made available on a continual basis to data science projects performing complex analyses in the data pool.

- Streaming or high-velocity data inbound to the logical data warehouse needs to be both [a] processed in real-time by machine learning or complex event processing (CEP) engines and [b] persisted historically for longitudinal analyses carried out by people, or programs, or both.

Breeding Success

Successful BDA teams, in their interaction with their internal customers, all exhibit certain fundamental, common behaviors:

- Under-promising and over-delivering: successful BDA teams set expectations with their customers in the business that they know they can exceed, communicate continually with business beneficiaries about those expectations while they are being met and exceeded.

- Continual involvement: successful BDA teams use their big data automation methods and tools to involve thought leaders in the business in the design and development process, showing those thought leaders what is possible — what success looks like — early, and often.

- After-the-fact story-telling: Successful BDA teams conduct thorough project postmortems with their business customers that emphasize (and demonstrate) not only value-delivered, but also the time-to-value-delivered, and the role of the IT organization's big data automation focus in producing the customer perceived value delivered. These post mortems are often cast as training sessions, and conducted in informal settings.

- Emphasizing systemic improvements in delivery capabilities: IT leadership's project post-mortem with its peers in the business, in successful BDA-enabled organizations, always emphasize the theme of systemic improvement in high-value delivery capability, through BDA — often by comparing past project time and value cycles to current time and value cycles.

Increasing BDA's Penetration within IT, Over Time

Experience tells us that, once we've set an expectation within the business as to the speed, flexibility and value of our delivery, the business will quickly assume that we'll always deliver value fast and effectively, and will increase the level of their demands.

That means that the big data automation mindset, and an organization's BDA platform, has to penetrate every area of the IT organization's big data practices, via continual use within projects, over a fairly short period of time — typically 12 to 24 months (1 to 2 budget cycles).

WhereScape®

Replacing Artistry with Automation

WhereScape's heritage as a supplier of automated data integration (ADI) environments dates back to the earliest period of commercial data warehousing, in the later 1990s. Along with our customers — hundreds of them, in 50 countries globally — we've lived through the slow and painful transition from hand-crafted, brittle and expensive data warehouses and data marts, to the age of data warehouse automation, and seen too many of our customers embrace too little automation, too late.

As big data technologies become an increasingly important part of our enterprise decision support environments — whether as staging areas, data engineering environments, sandboxes for data scientists and advanced analytics projects, or as a data management platform co-equal with our enterprise data warehouses — we should avoid making the same mistakes we made, with enterprise data warehousing, and instead build big data automation into our environment from the outset.

That big data automation infrastructure cannot, we think, come from the open source community or the Hadoop distribution vendors. Those companies have no real experience with, or understanding of, the existing enterprise data warehousing environment, or commercial decision-making in general.

They embrace — even promote — the very behaviors that got commercial IT organizations into such difficulties in enterprise data warehousing, by encouraging the broad-based, large-scale use of hand-coding techniques, in proprietary languages, using complex, poorly-integrated tool chains, with little or no support for operations and management of production big data infrastructure, and no understanding of the essential role that rich, well-managed metadata plays into the effective operation, and modification, of production-grade analytics environments.

In today's logical data warehousing environments, where relational data warehouses and data marts are integrated with, and cooperate with, big data infrastructure and proprietary data warehousing appliances, the automated data integration (ADI) tools we use to do a lot more, with a lot less, have to be equally fluent with, and fluid in, all of those environments, and apply common methods and practices across that heterogeneous infrastructure that ensures organizations do not, once again, become the victim of artisanal development practices.

That is WhereScape's mission. Big. Data. Warehouses. Right. Now.

About WhereScape

The pioneer in data warehouse automation software, WhereScape empowers organizations constrained by time, money or lack of resources, to deliver business value from their decision support infrastructure – including enterprise data warehouses, business facing data marts, and big data solutions. WhereScape has global operations in the USA, UK, Singapore, and New Zealand. www.wherescape.com