Wikibon.org

Follow the Money: Big Data ROI and Inline Analytics

David Floyer

A Wikibon Reprint

View the live research note on Wikibon.

Executive Summary

Earlier Wikibon research showed that the ROI on Big Data projects was 55¢ for each $1 spent. During 2014, Wikibon focused on in-depth interviews with organizations that had achieved Big Data success and high rates of returns. These interviews determined an important generality: that Big Data winners focused on operationalizing and automating their Big Data projects. They used Inline Analytics to drive algorithms that directly connected to and facilitated automatic change in the operational systems-of-record. These algorithms were usually developed and supported by data tables derived using Deep Data Analytics from Big Data Hadoop systems and/or data warehouses. Instead of focusing on enlightening the few with pretty historical graphs, successful players focused on changing the operational systems for everybody and managed the feedback and improvement process from the company as a whole.

The technical requirements of Inline Analytic systems are to enable real-time decisions within the current or new operational systems, without the current ETL (Extract, Transform & Load) processes that take hours, days or weeks to migrate operational and other data sources to data warehouse(s) and/or Hadoop systems. The software solutions developed by the winners deployed some or all of many advanced techniques, including parallelism, data-in-memory techniques and high-speed flash storage. Wikibon has researched different Inline Analytics technology approaches, including Aerospike, IBM BLU, Oracle 12c in-memory option and SAP HANA.

The key finding of this research is that the sponsors of Big Data projects should measure success by:

- The time taken to create Inline Analytic algorithms to automate decision-making directly into the operational systems of record, and,
- The effectiveness of supporting Deep Data Analytics to reduce the cycle time for improving the Inline Analytic algorithms and finding new data streams to drive them.

Where Big Data Makes Money

Figure 1 shows the results of a September 2013 study entitled "Enterprises Struggling to Derive Maximum Value from Big Data", where Wikibon found that the ROI on Big Data projects was 55¢ for each $1 spent. Management hoped for much higher long-term returns, with an average of $3:50 over a 3-5 year timescale.

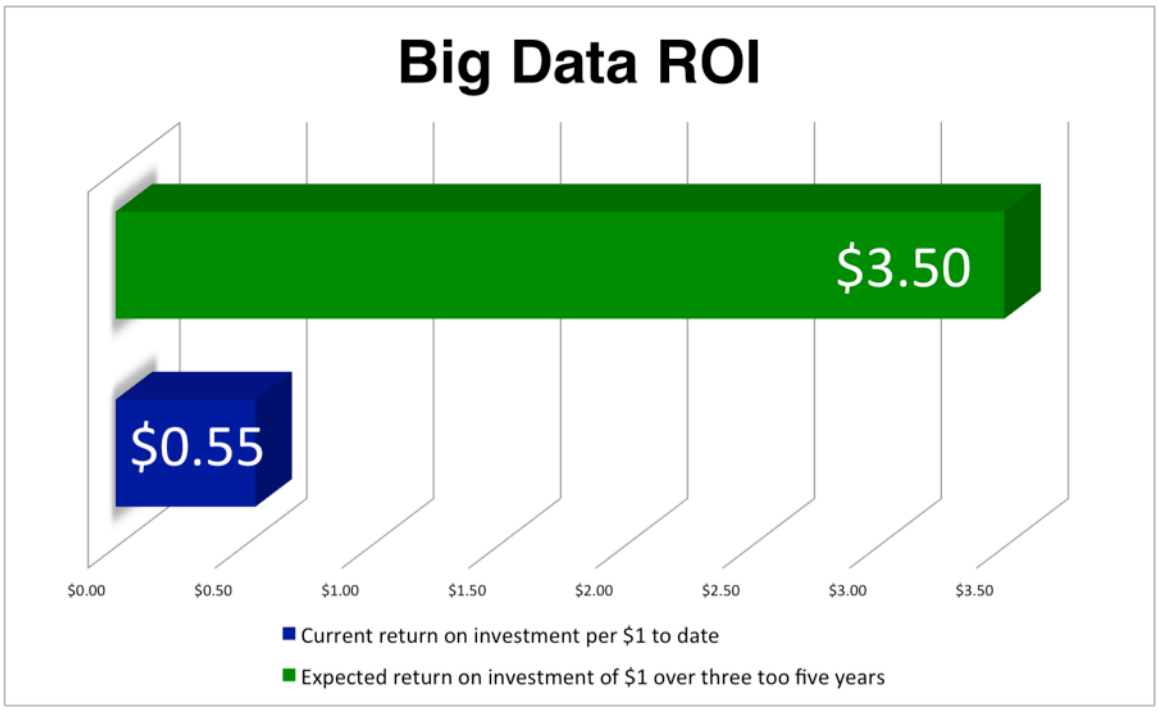

Figure 1 - Big Data ROI 2013
Source: Wikibon 2013,"Enterprises Struggling to Derive Maximum Value from Big Data", Figure 1

In this research, Wikibon focuses on specific organizations making great returns. Wikibon found a common thread in the way all of these organizations had implemented solutions:

1. Organizations had "operationalized" the Big Data processes and findings into current operational systems:
    - Algorithms had been derived that drove operational systems to improve organizational processes;
    - These algorithms were "inline", driven by real-time Inline Analytics, and integrated into operational systems;
    - Often, (but not always) the Inline Analytics were supported by tables derived from more traditional Big Data and data warehouse processes, the Deep Data Analytics part of the overall Big Data process.
2. The operationalization of the processes has radically changed the way Big Data is monetized and managed:
    - The traditional deployment of Big Data armed a few people in the organization with insight into issues and relied on them to make better decisions in the future (driving using the rear-view mirror);
    - The deployment of Inline Analytics focused on algorithms that used the data to directly change the way operational systems worked (driving with real-time input through the front windscreen);

- A small team of operational experts and data scientists use Deep Data Analytics to improve the algorithms, and to improve the quality of real-time data sources needed to drive the algorithms;
- The extended use of flash and data-in-memory technologies and techniques to avoid magnetic media bottlenecks are an essential component for the adoption of Inline Analytics.

Challenges for Big Data & Data Warehouse Processes

The traditional way of moving data for data warehouses and data marts is the **ETL process**:

- **E**xtract the data from multiple sources;
- **T**ransform the data to a common format, and reconciling differences;
- **L**oad the data into data warehouses tables and combine with historic data;
- Run stand-alone reports against the data warehouses;
- Create data marts and load into analytic tools for running stand-alone reports and for direct use by data analysts and data scientists.

The data movement rate for ETL processes is measured in terabytes/hour. The current maximum is about 5 TB/hour:

- The ETL elapsed times are constrained by the data transfer rates and "Amdahl's law", which points out that there are inevitably many single thread processes in a complex process such as ETL, and the elapsed time to completion is dominated by the compute and IO speed to complete the single thread components;
- Current systems are a trade-off between the amount of data loaded and time-to-value.

Often Hadoop is used as a cheaper and faster way of enabling the ETL process. When Hadoop is used this way, the elapsed times for ETL can be reduced. The costs of the ETL process can be reduced significantly by utilizing open-source tools and databases. However, ETL cannot become a real-time process.

The decrease of costs and elapsed time to do Deep Data Analytics is an important ingredient of gaining strategic and tactical insights, developing the algorithms to automate operational decisions and identifying potential ways to continually improve the algorithms, the data to drive the algorithms and the business processes to manage the technologies and skills required for success. However, Wikibon found that organizations with positive returns on Big Data projects have operationalized Big Data findings with Inline Analytics.

Technology Requirements for Inline Analytics

The technology requirement for Inline Analytics is simple to define and difficult to implement. The Inline Analytics must be a seamless part of the operational system and add no more than between 0.1 and 1 second to the total response time as perceived by the end-user, and retain the reliability, durability, consistency, recoverability and auditability requirements of transactional database computing. The Inline Analytics prerequisites include:

- In-Memory Databases, where the tables and key information is held in DRAM memory;
- The DRAM memory is protected against power failure by battery or capacitance technologies, and/or uses high-speed flash to protect and recover/restore data;
- All other data is held on high-performance flash technology (any traditional disk or hybrid disk technology would be a severe bottleneck to performance);
- The processing is highly parallelized, with high-bandwidth low-latency interconnects between processors, memory and flash technologies where necessary;
- All metadata about data is held in DRAM;
- Anticipatory fetching and processing enables faster access to supporting data from multiple data streams;
- Logical sharing of single copies of data (see Wikibon research "Evolution of All-Flash Array Architectures" for the impact of data sharing on storage costs) is built in.

All of these technology components are available to system architects. The most important single component is the in-memory database that will utilize the technology.

Technology Alternatives for Inline Analytics

In the sample of enterprises Interviewed by Wikibon and that successfully deployed Inline Analytics, the following database In-memory databases were deployed or tested:

1. Aerospike:
    - Aerospike is a flash-optimized in-memory open source NoSQL key-value database.
    - Aerospike handles very high volume streams of data, such as ad-serving applications required to bid within 100ms for the opportunity to place an advert following a web "click".

    - Enterprises with existing transaction applications would need to migrate them to Aerospike for integration with Inline Analytics.
    - Aerospike can be deployed as a high-performance analytics offload engine for separate data streams (e.g., Internet-of-Things data).
2. IBM BLU:
    - This is based on standard IBM DB2 OLTP 10.5.
    - "BLU Acceleration" is primarily designed for "read-mostly" Inline Analytics.
    - It utilizes and optimizes data in L1, L2 & L3 processor caches, and data-in-DRAM, but is not limited in total size to data that fits into DRAM.
    - BLU Acceleration includes columnar compression and allows some comparative analysis without decompression of data (saving on CPU register and cache resources)
    - Shadow tables uses the DB2 optimizer to determine the optimum use of either row-based or columnar-based schema at runtime.
    - Data skipping detect ranges of column values that are not needed to satisfy a query and avoids loading the data.
    - It can utilize SIMD (single instruction multiple data) instructions on IBM Power 7 or 8 chips to improve performance.
    - IBM has enabled the Oracle PL/SQL APIs in DB2, [1] allowing simpler migration from Oracle to DB2.
3. Oracle:
    - Oracle 12c is the latest version of Oracle, which includes an in-memory database option.
    - In-memory option has a list price of $23,000 per processor core, with an additional 18%/year of list price for maintenance and support.
    - Oracle is the most used enterprise database, with extremely high function and software high-availability options.
    - The traditional Oracle OLTP applications are row-based and read-write.
    - Oracle provides a columnar read-only in-memory portion for performing analytic queries in parallel.
    - It integrates with high-availability options such as Oracle RAC and Dataguard.
    - Oracle is also a supplier of many enterprise packages that use the Oracle database.
    - Oracle is a supplier of "Red Stack" hardware and software products such as Exacta and Exalytics that claim to provide a complete solution package.
4. SAP HANA (see link to Wikibon research "Primer on SAP HANA" for extended assessment):
    - HANA stands for High-performance ANalytic Appliance.
    - HANA is a pure in-memory design (any access to a table causes the complete table to be read and maintained in memory).
    - All the data for a HANA query must reside in DRAM.

- Background HANA processes maintain log files and migration to storage.
- SAP claims that HANA's architecture is designed to handle both high transaction rates and complex query processing on the same platform.
- SAP currently markets SAP HANA as in-memory analytics engine only appliance through several vendors including Dell, HP and IBM (OLTP & Analytics are not combined).
- SAP and its partners focus on the business value of faster reports from HANA appliances, which require to be loaded from a traditional ETL process.
- Currently HANA is not a Big Data analytics tool, or suitable for Inline Analytics.

There are many other in-memory databases, including Microsoft SQL Server 2014 with In-Memory OLTP extensions. See the linked Wikibon resources for additional research.

In summary, Wikibon has the following observations about vendors' offerings for Inline Analytics:

- Aerospike has the capability to process the highest ingestion rates, and is suitable for analyzing separate web data streams and Internet-of-Things data in real-time. Aerospike would need to operate as an offload engine to existing operational systems for data streams that were independent of the operational systems.
- Based on the interviews with customers and analysis of alternatives, Wikibon believes that IBM BLU is the most mature, relatively low-cost and high-performing solution that can integrate OLTP row-based systems and columnar data for Inline Analytics. However, for many customers that would entail a migration to DB2 for production workloads. IBM claims a large number of customers have migrated from Oracle to DB2 using the DB2 Oracle PL/SQL APIs.
- Oracle is the 800lb gorilla of the operational database world, with a long history of successful database development and operational excellence. Oracle 12c has a number of excellent features, including Pluggable Databases allowing Database Consolidation and rapid spin-up/tear-down in development environments. The Oracle 12c in-memory database option is clearly a first step in an in-memory computing journey. It requires significant set-up by DBAs and significant DBA maintenance. It is lacking many of the features found in IBM BLU, which would limit the scope of analytics that can be performed. The hybrid columnar compression is limited and tied to Oracle storage hardware. Wikibon believes that Oracle must and will devote significant resources to enable Inline Analytics to be delivered functionally and with automated operations. Wikibon would recommend Oracle break the tie between Oracle hardware and hybrid columnar compression.

However, Wikibon and the customers it has interviewed do not believe that Oracle 12c is sufficiently developed as an in-memory platform to perform other than initial proof-of-concept work or traditional reports that are time-critical to the business.

- SAP is positioning HANA as a long-term solution for Inline Analytics for SAP implementations. At the moment SAP HANA offers only support for relatively small data marts that can be held completely in memory. SAP HANA's greatest value is in providing specific operational reports that previously took days to receive, in minutes. However, HANA is a separate platform from SAP ERP, and set up and maintenance involves significant costs.
- An alternative and lower cost lower risk first step to SAP HANA would be to identify the reports that are most critical to the business and integrate in-memory analytics to produce these time-critical reports directly from either Oracle 12c or IBM BLU technology.

The analysis of successful examples in the next section suggests that far greater returns can be generated by developing algorithms that allow the decision making directly by changes to the current operational systems.

Inline Analytics Approaches that Led to Positive ROIs

Wikibon interviewed a number of companies that were using Inline Analytics successfully. The following two case studies, which illustrate the power of Inline Analytics, are based on eXelate and a bottling company.

eXelate is a data and technology company that improves and automates digital marketing decisions. The eXelate technology helps digital media be more relevant for consumers and supports more effective and accurate ad targeting for marketers. eXelate reduces Big Data to actionable smart data and algorithms injected into the operational systems. It ingests huge amounts of click and other Big Data streams from websites and other sources, builds models describing what that data means and uses those models and algorithms to populate operational database tables that optimize bidding for advertising space in real time (<100ms). The consumer is shown more relevant advertisements, and the advertiser increases revenue.

eXelate processes 60 billion transactions per month for more than 200 publishers and marketers across multiple geographic regions. Traditional on-line processing methods and databases are simply far too expensive. eXelate had to design a completely different way of achieving high performance and high availability. A detailed description of eXelate's IT Architecture can found at this link. The transaction infrastructure supports 60 billion transactions per month and 2 terabytes of data per day. The data is processed in the nearest Aerospike real-time NoSQL database cluster, using key value pairs. The Aerospike database has a flash-

first architecture to support 50% of the IOs being writes, and the SSDs are the persistent storage layer for recovery, further analysis by Deep Data Analytics and compliance. The decision tables are held in memory in compressed columnar format, and the algorithms in the applications process the transactions against the decision tables to decide if and how to bid.

eXelate's proprietary prediction models and algorithms are mainly developed on Revolution R Enterprise software and IBM PureData System for Analytics. This Deep Data system uses an IBM Netezza "shared nothing" parallel database appliance. The data comes from the front-end data capture servers, from the data transaction systems and from other data sources. The models and tables are then loaded into Aerospike databases as required.

eXelate has been pragmatic about the technologies it uses and focuses them on achieving automated business results through sophisticated Inline Analytics. Figure 2 shows the integration of different systems that are needed to exploit streaming data, operational systems, Inline Analytics and Deep Data Analytics.

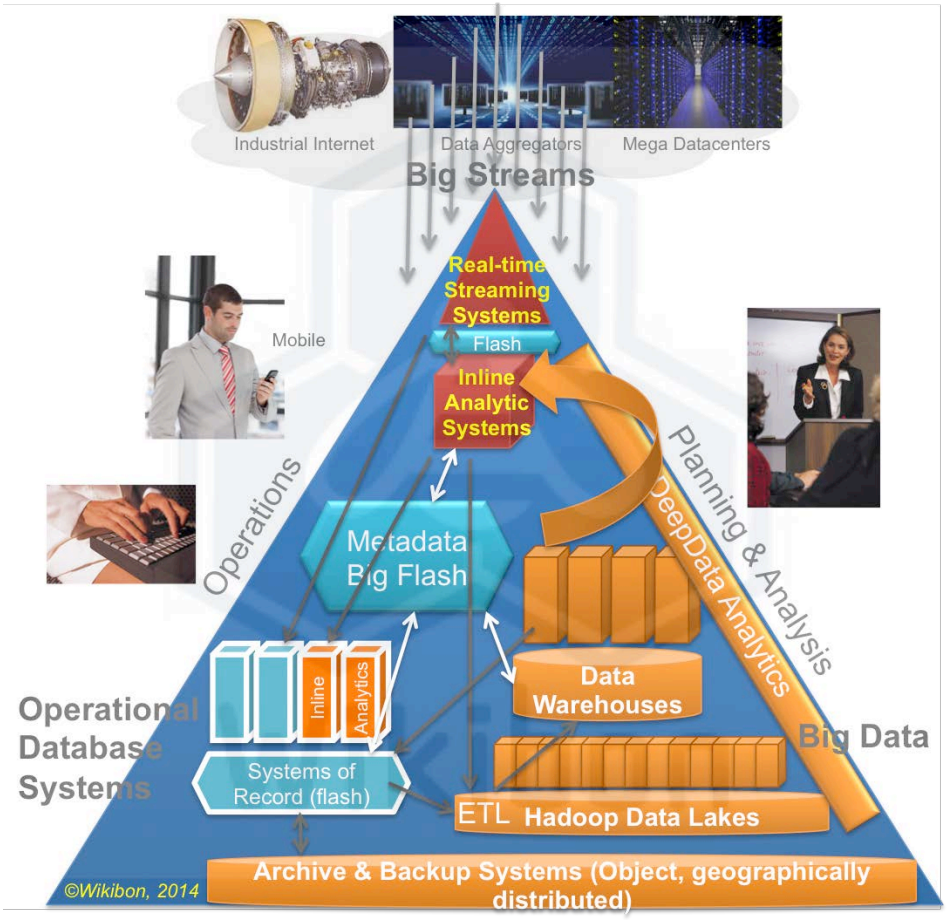

Figure 2 – Integration of Streaming Data, Operational Systems, Inline Analytics and Deep Data Analytics for Implementation of Enterprise-wide Integrated Applications

Source: © Wikibon 2014, based on Figure 6 in Flash and Hyperscale Changing Database and System Design Forever

Bottling Company (BC)

- BC makes, sells and distributes bottled products in multiple U.S. states. The company's operational systems use a number of SAP modules on DB2 to support financials, warehouse operations, materials management and customer data. The most important deadline is to be ready for the trucks to be loaded and have the drivers ready to roll first thing in the morning. Any failure during this process leads to a domino effect on the supply chain and deliveries and directly impacts revenue and profit. Problems also lead to traffic jams if the trucks are out of position and trouble with local authorities and the local community.
- BC has been using SAP on IBM DB2 for 5 years when it migrated from Oracle.
- BC joined the IBM BLU Acceleration Beta program.
- BC initially deployed BLU based on ease-of-deployment, data compression capabilities and query performance improvement.
    - With BLU, data compression now averages 90%; data loading times and compression have steadily improved during the beta process. Since it moved to IBM BLU with improved compression, BC has not had to buy any additional storage for its core database environments.
    - Query performance has improved dramatically with BLU Acceleration, especially for large analytic queries with a substantial calculation component (not so much improvement with simple tables).
    - The BLU added performance allows developers to test their code against production size large data sets; before they had to test against much smaller less representative data sets to avoid waiting for results.
    - The optimization features of BLU also make up for poorly written SQL queries.
- BC reports it has direct access to the team at IBM developing BLU, can ask them questions and get updates on the roadmap.
- The result of BLU is end-users get better performance and receive better data quicker. Drivers have all the information they need when they arrive first thing in the morning. The driver schedules are more efficient. The sales team has direct access to the real-time insights on iPads to help them close deals on the spot.
- Shadow table will allow BC to “simplify” its OLTP environment and better integrate the OLTP and Inline Analytics.
- BC evaluated the alternative of moving to SAP HANA and found it would require significant re-architecting and additional cost. BLU allows it the same performance as SAP HANA with better integration with OLTP and without the ETL delays.

- The business impact of BLU has started to show up directly in business results and peer benchmarking:
    - BC has cut its inventory on hand to one week, compared with the competitive benchmark of one month warehouse inventory.
    - BC had 5x better results last quarter than benchmark bottlers and turned good profits in a tough economic environment.

This bottling company has been able to provide Inline Analytics which led to improved ability to react automatically to change, and improved ability to provide the direct users (the drivers, warehouse staff and sales teams) with the information they needed when they needed it to do their jobs more efficiently.

Conclusions & Recommendations

Organizations that focus on translating the findings of Deep Data Analytics directly into Inline Analytics (i.e., develop automated algorithms and data sources directly into operational systems) will have much higher ROIs on Big Data investment. This focus will that will ensure that improvements will directly impact the organization, and will be the fastest way to drive continuous productivity improvements.

Action Item: CEOs and CIOs should measure success of Big Data by setting objectives about the time taken to create Inline Analytic algorithms to automate decision making directly into operational systems-of-record. The second measurement to focus on should be the effectiveness of ensuring that Deep Data Analytics focus on reduce the cycle time for improving the Inline Analytic algorithms, and finding new data streams to drive them.