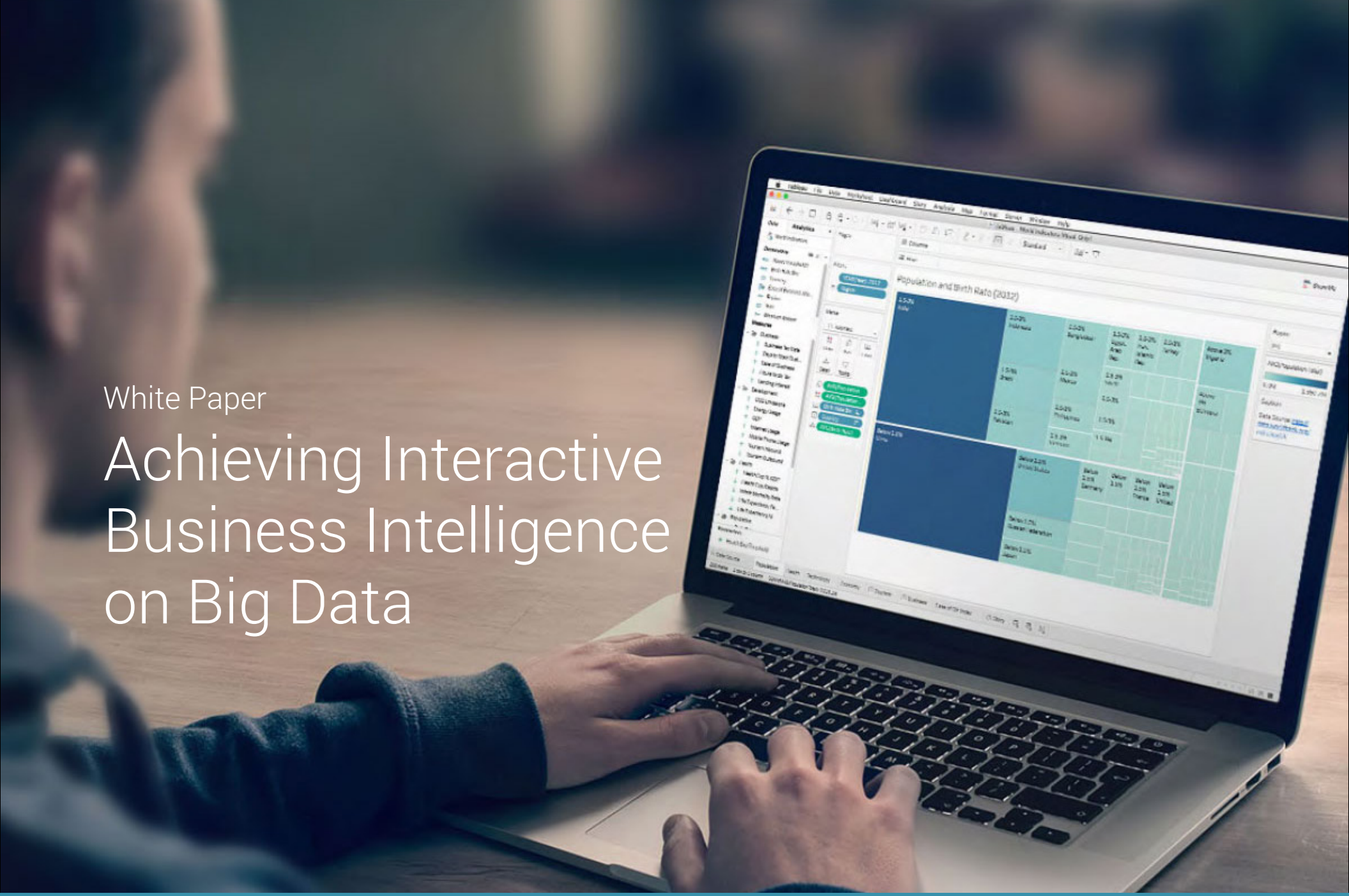

White Paper

Achieving Interactive Business Intelligence on Big Data

jethro

Introduction

Interactive Business Intelligence (BI) is a popular well-established method to analyze business performance. Interactive BI is used to find bottlenecks in e-commerce transaction streams, anomalies in profit and loss patterns of products and services, and to track the performance of a content delivery network by time and region—to name just a few applications. Interactive BI is popular across a wide spectrum of business segments, used daily by principals and managers at all levels.

The adoption of data driven decision-making unleashed a surge of business data and a rise in user demand to analyze it. This trend drives IT departments to migrate off their expensive Enterprise Data Warehouses (EDW) toward cost effective Big Data platforms like Hadoop or AWS. The new platforms present a financially luring proposition: a Total Cost of Ownership (TCO) per Terabyte that is about 10 times lower then the TCO of incumbent IBM, Teradata or Oracle EDW systems.

This dramatic cost savings proposition comes with a painful trade-off in tow. The new platform fails to match the high performance and user concurrency of legacy EDW systems for Interactive BI applications. With legacy EDW systems, users have come to expect response in seconds to complicated SQL queries executed over billions of records. In addition, the new platform is expected to serve thousands of concurrent users and keep the underlying data "fresh" to the last minute.

Business critical Interactive BI applications present a non-trivial performance challenges to Hadoop and AWS clusters alike. Enterprise IT is compelled to address these challenges before it can rip the dramatic cost savings offered by Big Data platforms.

Square Peg, Round Hole

Physically-coupled compute and storage cluster architecture delivers outstanding efficiency at scale. The clusters are designed to effectively combine the compute power, bandwidth, and storage of individual cluster nodes to serve a single task. The cluster relies on the following traits to effectively combine its resources:

1. Distributed – All cluster nodes are quasi-equal in compute power, storage, connectivity and therefore are interchangeable for the subtasks assigned to them.
2. Local Processing – Each cluster node closely couples compute power and local storage. This trait optimizes concurrent "share nothing" node level subtasks.
3. Redundant – Data blocks are systematically replicated across the cluster. Redundancy improves locality, but more importantly makes the cluster tolerant to failure of individual nodes. This trait simplifies adding and removing cluster nodes.

The architecture of the cluster is optimal for concurrent execution of a few "heavy" tasks. Distributed tasks, e.g. tasks that can be partitioned into multiple smaller quasi-equal node subtasks will benefit most from executing on a cluster.

The cluster is optimized to fully scan large-scale datasets at high throughput rates. Data Science workload such as Machine Learning, Predictive Modeling, and static reporting applications well suited preform well on Hadoop and AWS clusters.

Interactive BI workloads are notably different and are typically profiled by the following traits:

1. Concurrent – Multiple concurrent user sessions, possibly thousands, each firing a sequence of non- trivial SQL queries.
2. **Selective** – Interactive BI queries focus on significantly smaller subsets of the underlying data that pertain to the business area and timeframe analyzed. Each interaction may need to filter through billions of rows to find its unique subset.
3. Compute Intensive – Interactive BI queries typically aggregate numerical data like sales numbers and unit counts, across business factors like product, region, or month of the year.
4. Ad Hoc – Self-service BI apps are continuously changing with new queries being frequently added or modified. Underlying pre aggregated dada cannot be manually managed to track such changes.
5. Immediate – Interactive BI users expect results in seconds, so latency rather than throughput defines performance. Users expect the underlying data to be incrementally updated – continuously.

Query results are expected to be up-to-the-minute relevant. A typical interactive BI workload is far from fitting the optimal profile for a cluster and therefore mandates another layer to bridge the gap. Enterprises enjoy the TCO gains on their reporting, machine learning, and predictive modeling applications but suffer a costly struggle with their Interactive BI workload.

Jethro is a transparent middle tier that resides between the interactive BI clients and the underlying Hadoop or AWS cluster. Jethro is specifically designed to optimally utilize the compute power of a cluster to serve the most challenging Interactive BI workloads.

Cost of Interactive BI Service

Providing a solid interactive BI service is challenging enterprise for IT in more than one way. While business users revolt against any performance degradation or scale limitation of BI applications, enterprise IT is continuously pressed to reduce the cost to service them. To make matters worst, the lion share of that is the ever-rising cost of IT talent. Enterprise class Interactive BI has to address the cost of the service by the following:

1.Transparency Compatibility – Organizations commonly have tens, sometimes hundreds of interactive BI applications. Interactive BI acceleration solutions should transparently work with existing applications and avoid costly and risky modifications.

2. Automation – A BI acceleration solution must not require costly manual IT work to keep the data current and interactively accessible. Data re-engineering, index and cube management, and incremental updates should be automatic.

3. Adaptive – Interactive BI performance should be consistent and adapt to the growth of the underlying dataset and the changing nature of the application. Manual performance tuning is expensive.

In short – enterprise class BI is Interactive BI with a low cost of service.

Enterprise Class Interactive BI by Jethro

The Jethro BI engine is a transparent, self-driving, high performance interactive BI middle tier. It sits between the BI tool and the data lake. Users just point the application screen to Jethro to enjoy blazing fast response times; nothing else has to change.

Jethro drives itself – Indexes and cubes are created and maintained automatically to relieve IT of costly, error prone busywork.

Jethro uses several strategies to achieve high performance across a wide spectrum of queries:

1. Offload the "heavy lifting"

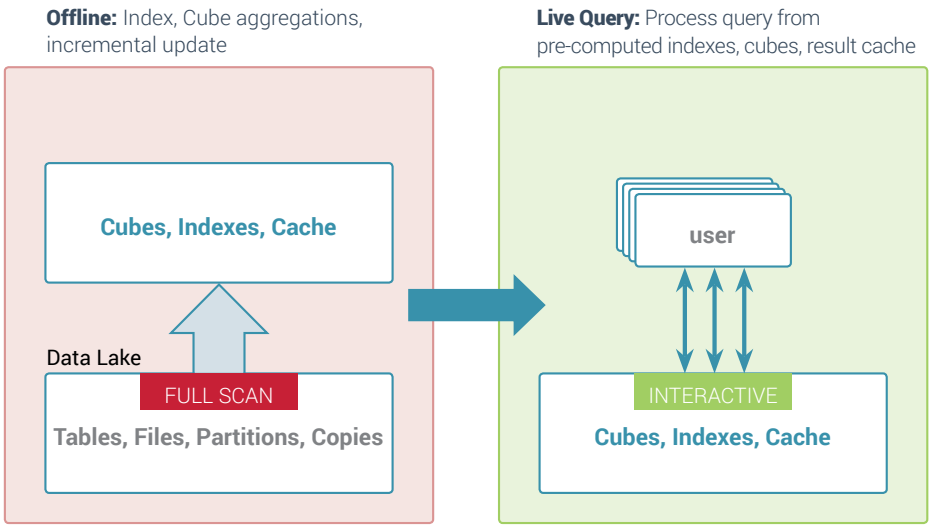

Figure 1 Jethro does the heavy lifting offline

Building and maintenance of cubes and indexes is offloaded to offline background processes. Query results are efficiently derived from the recomputed compact cubes and indexes rather than from the large underlying datasets.

2. Wide spectrum of queries

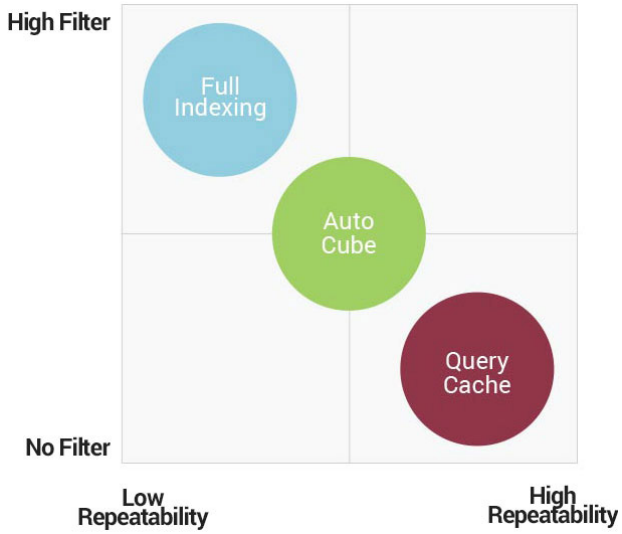

Figure 2: full indexing, auto cubes and query cache cover a wide spectrum of queries

Employing indexes, auto cubes, and a query cache delivers high performance on a wide range of queries.

3. Pre-computation – Think of indexes and cubes as partial computation of multiple future queries executed and stored ahead of time. User query results are derived from compact cubes and indexes, which requires significantly less computation.

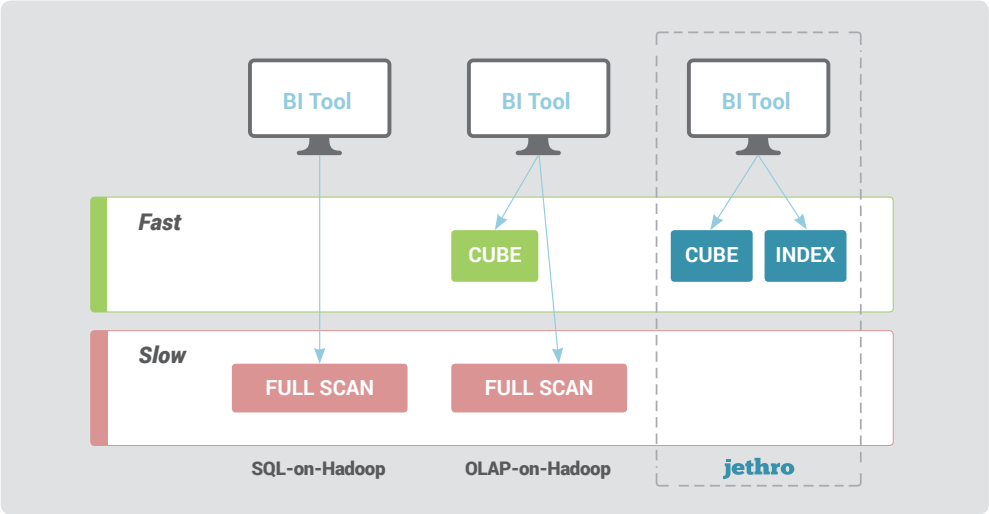

Figure 3: the Jethro advantage

Summary

Enterprise class interactive BI is a demanding use-case. In addition to a delivering high performance over large datasets to many concurrent users, the BI engine has to be self-driving. Enterprise BI customers demand high level of service while keeping a low cost of service delivery.

Jethro is designed from the ground up to do just that for on premise or cloud Hadoop clusters, and AWS virtual clusters.

Next steps: schedule a webex with one of our experts to help you run a POC.

About Jethro

Delivering Interactive Business Intelligence (BI) over Big Data is our passion and is our forté. Customers rely on Jethro to serve thousands of concurrent users analyzing tens of billions rows of data to support their business decisions. Actionable Business Intelligence mandates response time measured in seconds, up to date data measured in minutes, and data sets that span over 3 or more years of business. JethroData customers enjoy actionable, business critical BI at the scale, scope, and speed of their business.

Contact Us
+1 844-384-3844
info@jethro.io

Thanks for reading!

Let's chat and discuss how you can accelerate your BI to the speed of thought.

+1 (844) 384-3844
info@jethro.io