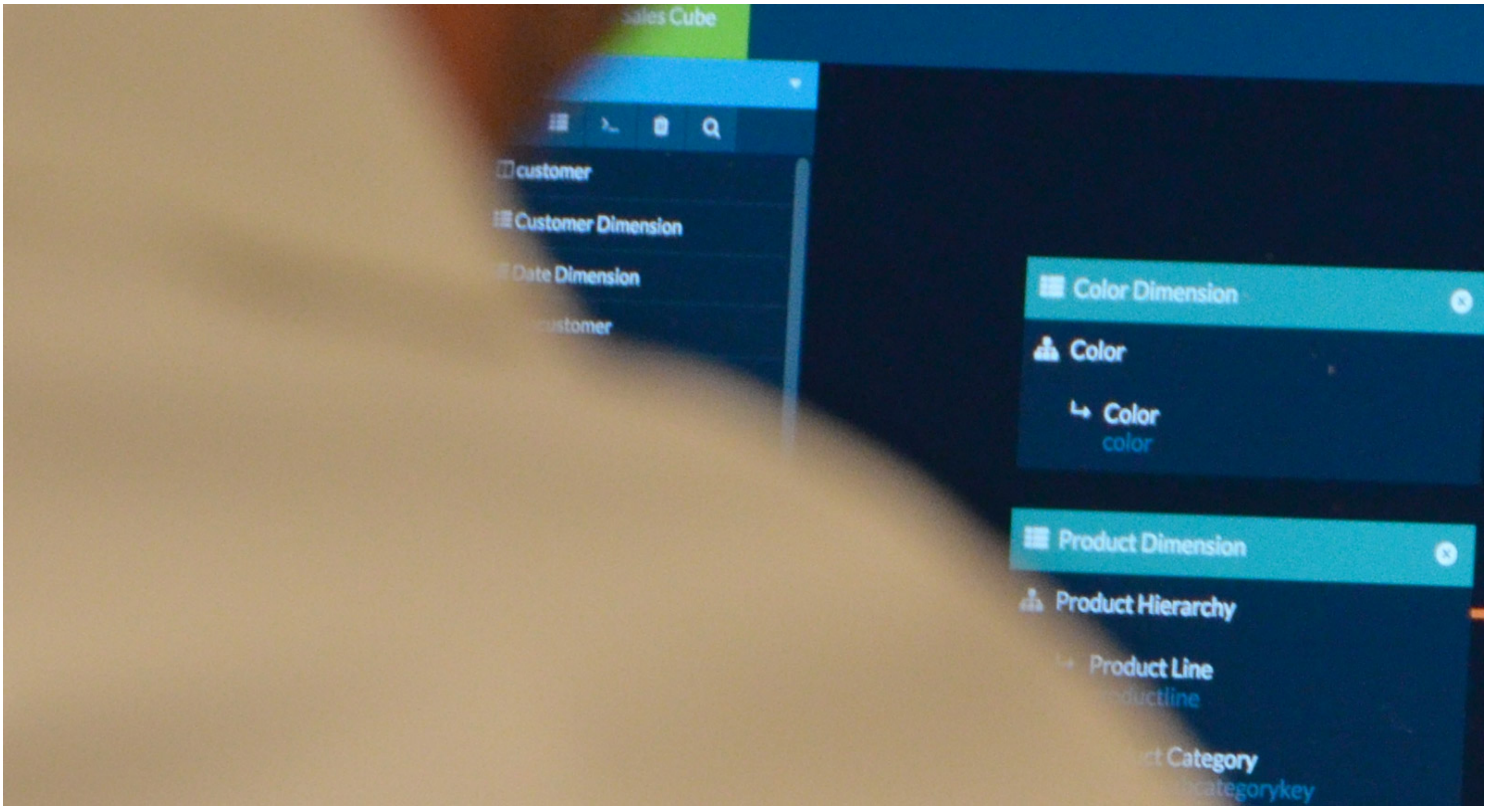
OLAP on Hadoop Solution Checklist

A Buyer's Guide to Evaluating OLAP on Hadoop Solutions



atscale

The Era of BI on Hadoop is Now

With Hadoop adoption on the rise, enterprises need to figure out how to bring their business intelligence capabilities to bear against massive data volumes, growing data diversity, and increasing information demands.

The goal of this document is to help enterprises understand how to evaluate solutions for doing interactive business intelligence on Hadoop-scale data sets. It covers what to consider when planning a new initiative for doing OLAP directly on data in Hadoop.

In the early days of Hadoop, data was only accessible to skilled developers and data scientists. Then purpose-built visualization platforms began to emerge to make the data in Hadoop visible to business users. While these new platforms provided basic visibility to the business, many of them failed to effectively make Hadoop work for business intelligence (BI). They were slow, inflexible, and required IT to move data into dedicated environments. Even worse, they required business users to adopt new visualization tools and data access languages in order to analyze data in Hadoop.

As enterprises moved more of their data into Hadoop clusters, SQL-on-Hadoop technologies emerged to query this data, and a new generation of OLAP-like solutions for Hadoop was born. As a result, the "BI on Hadoop" movement has matured tremendously. However, not all solutions are architected the same, and careful consideration needs to be made when determining the best solution for bringing BI workloads to Hadoop.

Installation & Configuration

Installation of the solution should be simple, prompted, and consistent. Upgrades should be seamless and require minimum downtime. Software components should work effectively with existing Hadoop ecosystem components.

Zero Footprint Install

Unlike other OLAP-on-Hadoop solutions, AtScale does not install any software on your Hadoop data nodes, nor does it require its own separate cluster resources.

Does the solution need to install software on the Hadoop cluster nodes?

Installing solution-specific daemons or software on Hadoop data nodes introduces cluster maintenance overhead, and may negatively interfere with other Hadoop components running on the cluster.

Does the solution require a separate cluster?

Hadoop is a data consolidation platform that was designed to support multi-tenancy. If you have invested in a Hadoop 'data lake', why deploy an additional cluster of hardware resources to do BI? Besides the additional cost, a dedicated BI cluster requires data movement and additional maintenance. Also, you will need to scale your BI hardware separately from your Hadoop hardware - exactly the type of activity Hadoop was designed to avoid.

Does the solution require additional Hadoop ecosystem components that your organization is unfamiliar with?

Many of the OLAP-on-Hadoop solutions require HBase for data storage. HBase is a very complicated application to manage and maintain. In addition, it's architecture requires frequent compaction of its data files, which results in severe performance degradation during compaction. Also, there are no good ways of backing up HBase data other than doing a full cluster replication.

Does the solution integrate with your Hadoop distribution's management console (i.e. Cloudera Manager)?

Being able to deploy and manage the solution with the Hadoop management console makes it easier to start and stop services and do client upgrades.

Do you have control over the solution's resource usage?

Hadoop is a multi-tenant architecture. The solution you choose should use a resource management pool like YARN so that its jobs can run simultaneously with other cluster workloads in a controlled fashion.

Does the solution have an installer and will it install all its services in one step?

Having to coordinate the installation of software on multiple nodes is error prone and time consuming.

Compatibility

AtScale was architected from the ground up to be compatible with virtually any visualization front end that is capable of speaking SQL or MDX. Additionally, AtScale doesn't create proprietary data structures or install custom code on your cluster.

Does the solution use a proprietary format for its data storage?

A proprietary data format is usually not compatible with other platforms in the Hadoop ecosystem including PIG, Hive, Oozie, HBase, etc. Proprietary formats mean data duplication and interoperability with other systems that use HDFS.

Does the solution use its own proprietary engine for processing data?

A proprietary, non-open source engine (unlike Hive, Spark SQL, Impala) locks you into that vendor's solution and doesn't leverage the innovation in the open source community.

Does the solution utilize it's own visualization interface?

A proprietary visualization interface introduces another BI tool variant that users need to learn. Most business users prefer to use the tools they already know, such as Excel and Tableau.

Does the solution require proprietary database drivers or client-side installation for data visualizations?

IT would rather not manage yet another client-side software installation component. This complicates wide-scale rollout and uptake.

Does the solution have a web-based design environment with multi-user support?

Since Hadoop is a multi-tenant environment - it must support multiple, simultaneous users accessing the data and designing OLAP cubes. Most legacy platforms limit model design and edit capability to a single user.

Does the solution integrate with external schedulers?

Updating model data and cube data should fit seamlessly into the enterprise's chosen scheduling/orchestration environment.

Does the solution fully embrace Hadoop's open source nature?

Utilizing as many open source components as possible makes your decisions more future-proof than relying on a single vendor's proprietary solutions.

Does the solution integrate with Hadoop's security systems (i.e. Kerberos, Sentry)?

Data security on your Hadoop cluster is a top priority, and integrating with the existing Hadoop standards ensures that the solution always shows the right data to the right people.

Does the solution integrate with LDAP for user/group management?

Onboarding users should be as easy as assigning them to a group in your LDAP or Active Directory environment. When scaling to 100s or 1000s of users, this becomes a critical requirement.

Maintenance

Analyzing data in Hadoop is not a "one-and-done" exercise. New data lands on the cluster every day, hour, or even minute. Querying the data in place and supporting automatic incremental updates make production BI workloads on Hadoop a reality.

Built for the Enteprise

AtScale's architecture was inspired by the operations of one of the world's largest data operations at Yahoo! The system has been built to handle the most demanding BI workloads.

Does the solution have a lengthy import process or make a separate copy of the data?

Creating data copies is expensive, introduces latency, and makes data synchronization difficult (if not impossible). Also, copying data may introduce data inconsistencies and make data lineage difficult to determine.

Can you easily upgrade the solution software in one step?

Updating multiple components/nodes for a software update is error prone and may cause prolonged system outages.

Does the solution allow for incremental data updates?

As new and updated data lands in Hadoop, the solution should be able to incrementally add or update just the changed data. Performing a full update on large data may not be feasible and introduces latency.

Does the solution require coordination/maintenance of another component (i.e. HBase) for the system to function?

Systems with several moving parts are more likely to fail and more difficult to maintain.

Is the visualization interface updated whenever the back-end is upgraded?

Retraining users is time consuming and costly and should not be dependent for server side (i.e. engine) only upgrades.

Performance & Scalability

Data in Hadoop isn't just bigger in terms of volume – it is also more complex. High-cardinality dimensions, unstructured data, multiple data formats are a struggle for traditional MOLAP approaches on Hadoop-scale data volumes.

How long before users can access new data in Hadoop? Does the solution have a real-time option?

Full cube builds or lengthy data update cycles limit system utilization. You should not have to compromise the data's scope or size in order to deliver on user latency requirements.

Does the platform return queries in 10 seconds or less, regardless of the size of the data set?

BI users expect queries to return in 10 seconds or less. Studies have shown that lack of interactivity causes users to abandon, cancel, or resubmit queries (causing system contention).

Does the solution handle a large number of dimensions?

Modern data sets often have hundreds if not thousands of dimensions. Solutions that use a MOLAP-style architecture cannot scale to Big Data dimensionality. In order to pre-calculate all data combinations, they have to either pre-aggregate the data or limit the data scope.

Does the solution handle high-cardinality dimensions (lots of distinct members)?

The solution must scale to support the size of dimensions in Big Data (millions of IP addresses, hundreds of millions of users, billions of web sessions).

Do distinct count queries perform the same as non-distinct count queries?

Distinct counts are very common in Big Data analytics. Large distinct queries can put an inordinate amount of load on the system due to large sort and shuffle operations, which can result in lengthy query times. Distinct count queries are not additive, meaning you can't use previously summarized data to compute the result. This makes it impossible to use aggregates to boost performance of these queries.

Does the solution handle at least 50 concurrent users?

The platform must handle many simultaneous user queries without significant performance degradation or without affecting other workloads on Hadoop.

Does the solution scale with your Hadoop cluster or does it require separate scalability planning?

System administrators should be able to scale their production Hadoop cluster without having to manage a separate set of resources. Administrators can focus on optimizing one cluster for peak load handling instead of incurring the time and cost of maintaining two clusters.

Does query performance scale with data size and Hadoop cluster size?

As the size of your data grows, you should be able to add more Hadoop nodes to scale query performance along with increased data size. The solution you choose should leverage this key advantage of Hadoop – horizontal scale out using commodity hardware.

Functionality

Enterprise-grade BI on Hadoop requires a rich set of functionality - standard language support for SQL and MDX clients, familiar multi-dimensional modeling workflows, and collaborative web-based cube design tools.

Does the solution support a cube metaphor (i.e. measures, dimensions, hierarchies, roll ups, crosstabs)?

Business users prefer an intuitive data interface, and the cube data model has proven to be extremely easy to use. The alternative is to require users to model data using SQL, which is too complicated for all but the most advanced users. In addition, this results in data inconsistencies as different business users create their own calculations (and version of reality).

Does the solution support drill-through to the most granular level of detail?

Business users like to start their data exploration at the most general level of summarization and drill down to explore the source of anomalies. If the platform pre-aggregates or transforms data, it makes it impossible to drill-through to the detailed view of data. Drill-through to the raw data is necessary to turn data exploration into actionable results.

Does the solution have an MDX interface that supports common MDX operators?

The MDX (multidimensional expressions) language is highly expressive and allows for the most sophisticated analysis functionality possible. Many BI tools, such as Excel, generate MDX queries.

Does the solution have a standard SQL interface?

Most BI tools and custom BI solutions use standard ODBC or JDBC drivers to access relational databases using SQL queries. The solution should be able to use existing drivers.

Does the solution work with your existing BI infrastructure?

Hadoop is the ideal platform for consolidating all of the data in an enterprise into a consistent, scalable data lake. The solution you choose should use Hadoop as the centralized data warehouse, and allow all users in the enterprise to access data using their current tools and processes. This eliminates data mart proliferation and numerous data copies.

Does the solution offer an intuitive, web based, multi-user design environment?

Having an easy and intuitive way to model data and define cubes is necessary for business user adoption. Business users should be able to self-serve their own data requests without IT intervention.

Do changes to the cube or model require a rebuild of the data?

Big Data systems often have billions or trillions of rows of data. Changes to the data model tend to be frequent, so long cube build times can impact downstream business processes, time to insight, and agility.

Does the solution designer allow for the reuse of components across multiple models?

The ability to reuse common dimensions, hierarchies and calculations across multiple cubes makes maintenance easier, faster, less error-prone, and more consistent.

(Continued on next page)

Functionality *(continued)*

Any Client, Any Time

AtScale supports your current business processes by allowing your analysts to use the BI tools they know and love. AtScale's virtual cubes allow data in Hadoop to be presented to the BI tools in a format they can work with. Users experience the interactivity and responsiveness that they have come to expect.

Does the solution natively handle modern data types or is ETL required?

Hadoop often captures data in its original application or log file state. This means data files in Hadoop are often denormalized and contain application-like data structures, such as maps and arrays. The platform needs to "unfold" these constructs "in place" or extensive ETL and data movement is required to make the data usable.

Does the solution support Hadoop's schema-on-read capabilities or is ETL and pre-aggregation required?

Solutions that require the data to be physically modeled, moved, and aggregated ahead of time reduce agility and create data silos. A modern OLAP-on-Hadoop solution should leverage Hadoop's ability to describe schema as metadata, and apply this logic at runtime - not rely on operationally complex ETL processing to prepare the data ahead of time.

Does the solution allow BI tools to query data live in Hadoop? Does it support Excel live pivot tables?

Live connections to the data allow users to access the latest data, and make changes to their models and cubes in real time, without having to wait for long cube builds. The most popular BI tool in the world, Microsoft Excel, uses a live connection for its pivot table reports.