Intel and Cloudera Integrate a Company's Oracle* Infrastructure with Hadoop* via Kafka* Messaging

Apache Kafka* messaging solution allows a financial service provider to integrate existing Oracle databases with a Cloudera enterprise data hub.

Why Intel and Cloudera

Intel and Cloudera take the guesswork out of Hadoop. Using a unique collaborative approach, we delivery excellent performance, security, and quality distribution, built on open standards. Working with more vendors across the ecosystem, a solution built on CDH can ensure freedom from lock-in, enabling you to build a robust big data solution to meet the needs of your business today and into the future.

- Uniquely aligned product roadmaps for software and hardware to drive innovation faster, providing many industry firsts with Hadoop.
- Deep partnerships with virtually every provider in the data center, streamlining the process for building Big Data solutions.
- Proven track records of identifying the driving industry standards, so you don't run the risk of stranding yourself on an island.

A leading provider of integrated financial services and institutional banking services in southeast Asia needs a solution that can easily and efficiently extend its data integration architectures to Big Data systems in real-time without negatively impacting the performance of their source systems.

The Company's data is provided both in near-real-time and in batch, for later analysis. After building a data ingestion pipeline to make data sources available to analytics teams, the Company can perform many business applications, and they will integrate many new data sources soon, ranging from OLTP databases to web server logs and message queues.

Results

The Intel/Cloudera solution yields the following benefits:

- Proven integration with Oracle GoldenGate* to ingest data from any Oracle database in the organization.
- No need for custom data dumps from Oracle databases, reducing costs.
- Built services for ingestion and real-time consumption, adding security and validation of messages.
- Automatic archiving of data to HDFS for later analysis in Hadoop.
- Automatic provisioning and continuous deployment.
- Starting to connect the first consumers of real-time and batch data for specific business goals.

Business drivers

Banks worldwide know they must improve their understanding of their customers by examining the volumes of data they currently possess about their customers and their banking habits. Many retail banks fear potential competition from large technology companies who have solved the Big Data equation and can siphon customers away with analytics offerings.

To that end, the Company wanted to provide event-driven feedback to its customers. To do so, they would need a faster response time than batch analysis could provide. At the same time, many data sources were not immediately accessible for analysis because of organizational silos or security constraints. Providing a platform for ingestion and simplifying the ingestion process would save the hassle of having to run customized projects.

The Company came to Intel for assistance configuring a Hadoop-based Big Data solution that would integrate seamlessly with their existing investment in Oracle solutions.

Solution details

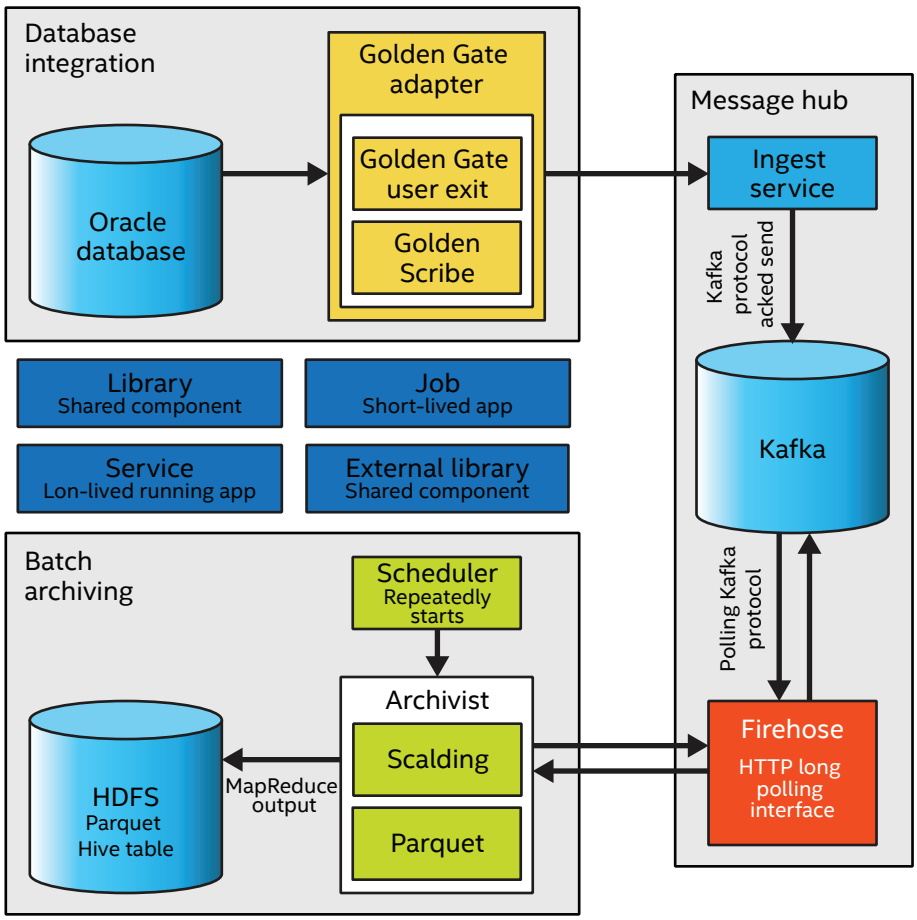
The solution (*Figure 1*) centers on Apache Kafka*, an open source publish-subscribe (pub-sub) message broker designed as a distributed system, which makes it easy to scale out. Kafka offers high throughput for both publishing and subscribing processes, and supports multiple subscribers. It is fast, scalable, and durable, and unlike traditional message queue systems, Kafka's messages require small overhead per message.

Because Kafka messages persist on disk, it is possible to read data from any offset in a log within a retention interval. Kafka is suitable for batched consumption and intermediate HDFS data storage as real-time events.

Kafka also provides the following benefits:

- **Data storage/ingestion.** Apache Kafka is used for intermediate storage of data. Oracle GoldenGate* was used to get the data from any OLTP Oracle database, since it requires no change to the existing database schemas and Oracle DBAs are familiar with it. Because HTTP is a well-known standard for many and is easy to use from any programming language, Intel built custom HTTP REST services to wrap Apache Kafka.
- **Security.** All data needs to be validated before it is stored in Apache Kafka, and deduplication of data is required (which can happen in certain failure scenarios). Because the Apache Kafka protocol currently does not support security mechanisms like authentication and authorization, we used Transport Layer Security (TLS) protocol cryptography to help secure the services and Unix pluggable authentication modules (PAM) for authorization.

Figure 1 **Kafka messaging in an Oracle GoldenGate* environment.** Existing Oracle databases pass data through GoldenGate to an Apache Kafka-based message hub, which shares information from the Cloudera enterprise data hub through PAM-encrypted HTTP messages.



- **Wire protocol.** All messages going in and out of the system are serialized in Apache Thrift* format. This format is already widely used in the organization, especially on Hadoop.
- **Data consumption.** Apache Kafka can handle millions of messages per second. This data is consumed directly from an HTTP fire hose service. An archiving service built in Scalding also connects to this fire hose just like any other client and writes data in Parquet format to HDFS for further processing. We built a fire hose command line client for developers so they could browse through the data in Apache Kafka topics. The solution will also have metadata service to make data discoverable.
- **Horizontal scalability.** Apache Kafka is built to scale out. Topics are partitioned across many servers.
- **Replay.** Apache Kafka can replay data from any point in a topic/stream, making it possible to process message streams in a fault-tolerant fashion.

The Company was happy with the integrated system and confident they could maintain the two systems side by side without major upheaval.

Cloudera Enterprise

The Company chose Cloudera for its ability to work seamlessly with existing Oracle databases and its cost-effectiveness, scalability, workload heterogeneity, enterprise-wide features, and open source framework.

Storing and processing historical data in Hadoop is very cost effective comparing to traditional data warehouse solutions. The Hadoop platform is an inherently scale-out architecture, which makes it easy to add nodes as web data volume increases. Because the Company expects data streams to grow exponentially as their business environment changes, the cost of adding storage becomes predictable.

Cloudera provides massively parallel fault-tolerant processing without the need for complex application-level coding.

Cloudera Enterprise lets customers handle rapidly increasing volumes of data and a variety of workloads from existing systems while optimizing the efficiency of such legacy infrastructure.

With the flexibility to run a variety of enterprise workloads—including batch processing, interactive SQL, enterprise search, and advanced analytics—to support their diverse data-driven goals, the Customer also benefits from Cloudera's unified view of information.

At the enterprise level, Cloudera offers several key features needed for IT compliance, including encryption at rest and in motion, Simple Network Management Protocol (SNMP) support and alerts, rolling updates, AD/Kerberos integration, and automatic backup and disaster recovery (BADR).

And lastly, because Cloudera distribution of Hadoop is based on open source components, the Company can integrate CDH with other open source tools, such as Apache Kafka, which is included with CDH.

Summary

The Company can now offer its customers business insights from similar industries in comparable locations in the form of real-time data analysis on their business's performance relative to competitors. This kind of Big Data-driven competitive analysis is something most small businesses cannot afford to create for themselves but find very valuable.

Intel was able to design a system that helped the Company integrate its existing Oracle infrastructure with a scalable Hadoop-based Big Data platform and provide additional services like these to customers.

Let us help your business too.

Spotlight on Cloudera

Cloudera is revolutionizing enterprise data management by offering a unified platform for Big Data, an enterprise data hub built on Apache Hadoop*. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Cloudera's open source Big Data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,800 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

For more information, visit **www.cloudera.com**.

cloudera®

Meeting your needs

We look forward to meeting with you to define your requirements and meet your objectives.

- **Accelerate time to value:** Achieve real-time cost savings, respond to market trends, and drive innovation.
- **Secure Big Data:** Deploy a sustainable Big Data program that doesn't put your organization, or you, at risk.
- **Maintain control:** Work with a partner who educates your team so you become self-sufficient.
- **Increase business potential**: Create and execute a plan that helps you adapt now, and in the future.

Hadoop sizing guide

		Cluster size		
		Small	**Medium**	**Large**
CPU		**Intel® Xeon® Processor E5 v3**		
Storage (TB)		<72 TB	72 to 570 TB	>570 TB
Node count	**Master**	2 to 3	4 to 7	≥8
	Slaves	<12	12 to 95	≥ 96
Memory (GB)	**Master**	64 GB	128 GB	≥256 GB
	Slaves	48 GB	96 GB	≥128 GB
Network		1 Gbps	10 Gbps	10 Gbps

Hardware configuration is highly dependent on workload. A high storage density cluster may be configured with a 4 TB JBOD hard disk, while a compute intensive cluster may be configured with a higher memory configuration.

Contact us

Contact your sales rep or e-mail us.

Intel.com/bigdata/services

333170-001