Prime Analytics Overview

Cisco Prime Analytics is a real-time big data analytics platform that allows you to connect to continuos streams of NetFlow, syslog, and XMPP data and run queries on both live and stored data to answer real time business questions.

The following topics provide an overview to Cisco Prime Analytics features and architecture:

Features and Functions

Prime Analytics provides the following features and functions:

- Streaming data—Prime Analytics can process large streaming data volumes, which allows decisions to be made based upon current data trends and analysis.
- Continuous queries—To interpret continuous data streams, continuous queries are required. Prime Analytics continuous queries are always on. New results are generated whenever new data arrives. The data is routed through the active query set and new results are published to downstream subscribers.
- Windows—Prime Analytics windows allow continuous data streams to be divided into segments so that results can be presented at meaningful points. Windows can be based on time intervals or by the number of arriving records.
- Streaming views—Prime Analytics allows you to create views of streaming data for higher-level real-time data analysis.
- SQL support—Prime Analytics supports SQL for real-time data stream queries. Supported SQL functions include stored procedures, user-defined functions, user-defined aggregates, joins, and other functions.
- Replay and drill-down—To learn from a continuous event stream and improve organization responses to them, Prime Analytics allows you to replay, analyze, and drill into data. This allows you to evaluate and tune organization event responses.
- Shared processing—To handle continuous query performance demands, Prime Analytics uses parallel processing to maximum advantage and ensure performance degradation does not occur.
- Adaptability—The Prime Analytics continuous query engine is adaptive; it can accommodate dynamic additions, removals, or modification of queries on-the-fly, without requiring a system restart.

- Support for data streams and data tables—Prime Analytics allows you to run continuous queries over data streams and data tables.

Architecture Overview

Prime Analytics components include a continuous data analysis engine that handles both real-time and historical data, a data integration layer that moves data into and out of the system, and a visualization layer that provides dynamic dashboards and reports to end users.

Prime Analytics components include the continuous query (CQ) engine, integration framework, and business intelligence (BI) platform.

CQ Engine

The Prime Analytics continuous query engine, called TruCQ performs all continued and historical data processing functions and stores results in the PostgresSQL Prime Analytics database.

Integration Framework

The Prime Analytics data integration framework is a Java platform that allows you to configure connectors and handlers to external data sources and data targets. The integration framework customizes communication with the CQ engine at the data input and output points. Integration framework connectors handle data transport protocol integration. Connectors are drivers that connect to external applications and support the input of the streaming, batch, structured, and unstructured data types. Connectors also handle data transformation functions as required by the query engine or the target application, for example, reformatting dates or combining a device ID with static data contained in an external database.

Any given data integration point involves multiple aspects including transport, format, schema, and others. With data input, for example, transport governs how the data arrives at the integration framework and is generally associated with a wire protocol such as HTTP, NetFlow, syslog, and XMPP. Format governs the message payload. It can take various forms such as delimited text, XML, JSON, Google Protocol Buffers, Apache Thrift, and others. Finally schema governs the integration point output. It is a structured format suited for pushing data into the continuous query database.

To handle the transport, format, and schema requirements, the Prime Analytics integration framework offers two mechanisms: connectors and handler chains. A connector is a single monolithic Java component that is responsible for all data integration point areas. You could, therefore, have a single component that extracts XML data from a JMS queue and massages the XML payload according to an Extensible Stylesheet Language stylesheet to render the output as a tab-delimited structured schema.

In contrast, the handler chain modularizes the various connector elements into independent handlers that can in turn be chained together to solve a given data integration problem. The handler chain could, therefore, have separate handlers for extracting arbitrary objects from a JMS queue and for converting an XML object using an XSLT stylesheet. The primary advantage of the handler chain is reusability. If, for example, you have a new data integration point that sends JSON data over a JMS queue, or XML data over an XMPP transport, you simply need to implement a single new handler for each case.

Connectors and handler chains are both supported, However, the handler approach is recommended. An additional advantage of the handler chain is the ability to take a user-implemented handler chain and turn on multi-threading.

BI Platform

The Prime Analytics BI platform is built from third party open source components including Pentaho, Saiku, and others. The BI provides the visual representation for continuous query data and provides tools you can use to create dashboards and reports that present data in ways suited to your specific business needs.

Continuous Query Engine

The Prime Analytics continuous query engine is the foundation of the Prime Analytics architecture. The query engine performs all data processing and analysis functions, including standard relational database operations as well as continuous queries against continuous streams of data. A data stream is an unbounded, potentially infinite, series of records, or tuples, traveling through a network. Similar to data tables, a stream is a database object with an associated structured schema. Data stream examples include:

- Flow events from network routers and switches.
- Video player logs.
- Network event or security logs.
- Website click and impression stream data.
- Sales data from distributed point-of-sale terminals.
- Data feeds from sensors, barcodes, or radio frequency identifiers (RFIDs).
- Financial tick data from securities exchanges.
- Transaction data over debit or credit card networks.
- Service-oriented architecture (SOA) components on an enterprise message bus.

A data stream is any series of tuples that grow over time. The decision to categorize and manage these data flows as streams instead of writing them to database tables is based on the data volume, sequence, and the latency requirements of the applications that depend on the data.

Prime Analytics streams can be one of the following types, depending how the stream is populated:

- Raw streams—Are populated by an external data provider that connects to Prime Analytics using a well-defined protocol and pumps in data for the Prime Analytics to use.
- Derived streams—Are defined using a query, called a defining query, and populated by the continuous query engine. A derived stream can be one of two types:
  - View—Similar to a regular database view, a streaming view creates a virtual stream that can be used by other queries instead of a raw stream. A view's defining query only runs when a query using the view is running.
  - Persistent continuous query—Is a materialized continuous query explicitly associated with a stream. It is similar to a view, but does not have macro semantics and is always active.

Each stream record that enters Prime Analytics has a time stamp attribute. (This attribute is also called cqtime.) The time stamp attribute is either provided by the source or user, or is system-generated. The user-provided time stamp is preferred when the original event time stamp is required to establish a continuous query window. The system-generated time stamp is preferred when the original time stamp is not important.

Windows

Prime Analytics uses windows to divide the continuous data streams into discrete data sets. Certain operations, such as aggregation, require finite data sets to generate results. Windows may be time-based or record-based, depending on the query requirements. Windows are computed using time-stamp attributes for time-based streams and internal record counters for record-based windows. The window defines a snapshot for a given data stream to create a finite set of tuples on which queries can produce results.

Continuous Queries

Continuous data stream queries run continually and concurrently. They produce results in continuous, zero-latency, output streams. Prime Analytics queries follow SQL query syntax. Any SQL queries and data processing models created for traditional database systems can be used in Prime Analytics. Supported SQL query functions include:

- Views.
- Filters, for example, such operators as >, <, and =.
- User-defined functions, for example, C, C++, Java, and Perl.
- Joins over streams and relations
- Aggregates, for example, SUM, COUNT, MIN, MAX, and AVG.
- User-defined aggregates.
- Grouped aggregates.
- Arbitrary subselects.
- Other subqueries

The continuous query engine performs queries directly against the data streams. It does not need to persist data to generate results. However, you can archive queries so they are persisted to the analytics tables for later access, replay, drill down, analysis, data enrichment, or historical correlations.

Replay and Drill Down

During an ongoing flow of continuous stream queries, you sometimes might want to review an event sequence again to fully understand what happened. When this occurs, Prime Analytics allows you to pause and replay the stream to review the event sequence. You can also take a data snapshot and drill into it to analyze the event and the system response. This replay and drill down capability enables you to better understand how to respond to business events and tune responses over time. You can take a snapshot of data from a particular time sequence, replay the stream, then create drill-down reports to understand and learn from the event sequence.

Prime Analytics Stream Relational Database Management System

The Prime Analytics continuous query database is a relational database extended for doing big data analytics. The continuous query database is based on the PostgreSQL open source object-relational database system. PostgreSQL functionality is available for continuous query data including tables, caching, archiving, replay, queries over streams. PostgreSQL modules packaged with Prime Analytics include the postmaster, system catalog, query parser, query optimizer, ODBC and JDBC adapters, user-defined functions and aggregates, as well as plugins for external programming languages.

Integration Framework

The Prime Analytics data integration framework is a Java platform that allows you to configure connectors to external data sources and data targets. The integration framework communicates with the continuous query engine at the data input and output points over a unified data integration layer. Integration framework connectors handle data transport protocol integration. Connectors are software components that connect to external applications. Connectors support the input and output of streaming and batch data sets. The data sets can be naturally structured or unstructured. Connectors also handle data transformation functions as required by the query engine or the target application, for example, reformatting dates or combining device ID with static data contained in an external database.

Connectors can be used for either input or output of data. Input connectors, also called producers, integrate external data sources that send data streams to the query engine. These include message queues, enterprise buses, proprietary APIs, and other data sources. Input connectors support the following source types:

- Pull sources—Used for traditional database systems.
- Push sources—Initiated by the connector (push-client) or by the data source (push-server).

Output connectors, also called consumers, deliver query results from the query engine to external source systems. Continuous streaming queries produce data streams. Like input connectors, output connectors can support different modalities depending on the client interaction:

- Push targets—Used for clients that can accept continually streamed query results.
- Pull targets—Used for clients that support only intermittent result retrievals.

Prime Analytics provides a metadata catalog of common connectors that cover most major data formats including:

- Syslog
- NetFlow
- XMPP
- CSV
- JDBC
- ODBC
- JMS
- Log4J
- File
- Retry
- Throttle

Connectors are created using an XML template. You can modify connectors or create new ones using an XML editor.

Business Intelligence Platform

The Prime Analytics business intelligence platform consists of open source applications developed by Pentaho, Saiku, and others, to create a web-based, client that can generate dynamic dashboards and reports from continuous query data. The BI platform includes functions for integration, information

delivery, and basic analysis. Information delivery elements include reports, dashboards, and ad hoc queries. Basic analysis elements include Online Analytical Processing (OLAP) and interactive visualizations.

Security

Prime Analytics security functions include:

- Role-based password-protected access for multiple users
- Multiple user authentication methods (PAM-based and standalone)
- Web-based and CLI-based user management
- Password enforcement policies (aging, minimum length, and lockouts)
- Audit trails of all user actions and all access through the web interface
- Security logs

In addition, you can enable SSL for added security between client web browsers and the Prime Analytics BI platform.