White Paper
Analytics and Big Data

Shine a Light on Dark Data with Vertica Flex Tables

Hidden within the dark recesses of your enterprise lurks dark data, information that exists but is forgotten, unused, and neglected. Dark data is rarely seen, but it contains a potential treasure trove of valuable information that can improve revenue, make your enterprise more efficient, or help you make breakthrough discoveries in your business. Interest in dark data is growing among corporations and data scientists as they strive to find additional intelligence within their data assets. The realization that dark data may contain crucial customer information, sentiment analysis, or advances in product development leads data scientists to explore dark data.

Table of Contents

This white paper describes the data exploration and data analysis functions of Flex Tables.

Big Data Challenges

While you may recognize the immense potential value in big data, you also may find it particularly challenging to explore certain forms of semi-structured data such as social media data, Web logs, sensor data, and other dark data sources. These critically important data types traditionally have required a lengthy process to load into legacy data analytic platforms and data warehouses before they deliver value. You want to explore, analyze, and shine a light on this dark data to exploit its potential.

Some of your challenges to managing this data include:

- **Extreme volumes**—since data volumes are so large in modern enterprise solutions, data should be stored on a scalable architecture that is easy to manage and minimizes data transformation and movement.
- **Dynamic schemas**—since the speed of innovation is accelerated today vs. legacy systems, modern data structures have data formats that change more often than traditional enterprise data systems. Modern architecture needs to handle today's dynamic data.
- **Licensing costs**—legacy architectures often have licensing models that challenge the cost-effectiveness of big data analytics, especially when it comes to data whose value is not yet proven. Modern architectures should offer value based on realistic data volumes of today.
- **Value of data for analytics**—Dark data value is often unknown. Modern analytics platforms should allow users to explore dark data freely to understand its value, without necessarily committing massive resources to it up front.

Vertica developed Vertica Flex Tables specifically with these big data-dark data challenges in mind. Flex Tables are designed to let data scientists explore dark data and fully exploit its potential. This white paper describes the data exploration and data analysis functions of Flex Tables.

Vertica Meets the Challenges

To understand Flex Tables, it's important that we first understand Vertica as a whole. Vertica is designed to manage large, fast-growing volumes of data and provide very fast query performance when used for data warehouses and other query-intensive applications. Vertica provides drastically improved data management and query performance over traditional relational database systems, provides high-availability, and requires no special hardware to achieve these results.

Its design includes:

- **Column-oriented storage**—increases performance of analytical data analysis queries
- **Compression**—improves performance by lowering I/O bandwidth with lossless compression
- **Standard SQL interface**—provides a familiar way to access the data with no gaps in functionality
- **Advanced data management**—makes Vertica easy to use and maintain through automated data protection, server recovery, query optimization, and storage optimization
- **Support of hybrid infrastructure**—Uses the same powerful Vertica analytics on data that is on-premise, on the cloud, or in Hadoop

By integrating these less structured data sources and supporting vanilla SQL queries against them, we bring a key feature of relational databases to bear—abstracting the storage representation from the query semantics.

Vertica Flex Tables

Flex Tables enable Vertica to query unstructured data or the dark data that exists in your company. Vertica gives you the power to quickly and easily load, explore, and analyze semi-structured data, such as social media, sensor, log files, and machine data

With Flex Tables, you can explore and visualize information such as JSON and delimited data without burdening or needing to wait for your IT organizations to extract, structure, and load the data. Flex Tables remove the need for coding-intensive schemas to be defined or applied before the data is loaded for exploration. Flex Tables create data exploration schemas as needed for high-performance data analytics and deals with the ever-changing structure of data with greater ease. Vertica does this by deriving structure out of the current file, as long as the semi-structured data has the following characteristics:

- The data consists of many records representing discrete sets of information encoded in some semi-structured data format.
- Each record has a set of addressable information. This means that some key can refer to each piece of information, either context-sensitive or canonical. A canonical address or key would be "author" found in a JSON map.

A context-sensitive address would be "the second timestamp" in a log line or "the third column" in a CSV file lacking a header row and could be assigned keys such as "Column 3."

There is also some flexibility in Flex Tables regarding anomalies in the data itself. You can't expect that all semi-structured data would necessarily be static and trouble-free. In fact, Flex Tables can handle situations such as:

- **Data variability**—records in a single set of unstructured data can vary their key space, structure, and information types. A single unstructured data set can have entirely unrelated records (e.g., records about books in the same data set as records about the history of forks).

Having a function like Flex Tables that automatically understands common schemas is powerful. It takes time to map structure and set up schemas, slowing the overall process of exploration of the data.

- **Schema variability**—Flex Tables allow for related records of variable schema. You may have data that, for example, has "zipCode" as a number, another record has it as a string, another may have a "locationZip," and others do not have it at all.
- **Nested objects**—the information of a record may be arranged in a hierarchy and have relationships with other information within the record. For example, JSON allows nested objects with a record.

By integrating these less structured data sources and supporting vanilla SQL queries against them, we bring a key feature of relational databases to bear—abstracting the storage representation from the query semantics.

How Flex Tables Differ from Native Vertica Storage

The capabilities offered in Flex Tables differ from native Vertica tables in a number of ways:

	Native Vertica	Flex Tables	Legacy Database or Data Warehouse Appliance
Column-oriented storage	X		
Compression	X		
Standard SQL interface	X	X	X
Advanced management	X	X	Limited
Analytics speed	Fastest (native Vertica)	Faster (Flex Tables)	Limited (Legacy)

Table 1. Feature comparison

Because native Vertica tables offer compression and columnar storage, they offer performance advantages over Flex Tables. That said Flex Tables are optimized for exploring semi-structured data. If you have unknown dark data sitting in storage and need to understand whether the data can deliver analytical value, use Flex Tables to perform data exploratory analytics. The expense and hassle to explore dark data using Flex Tables are low. Flex Tables provide an easy pathway for dark data exploration.

Both data analysis solutions offer huge advantages over legacy database solutions in terms of cost, speed, scalability, and manageability. Analytics that leverage the features of native columnar store have proven to be thousands of times faster than traditional relational databases, data warehouse technologies, and even Hadoop-based solutions..

How It Works

Let's take a look at how Vertica ingests data, applies schema, and stores data using its massively parallel processing (MPP) architecture when using Flex Tables.

Data Ingestion and Determining Schema

A major power feature of Flex Tables is its ability to readily ingest and apply schema to unstructured and semi-structured data. Common data types in the Big Data world aren't completely without structure. For example, JSON often contains XML tags that contain structure, even if that structure may not map readily to a relational schema. Delimited files often contain header rows and structure between delimiters. Vertica understands this structure (and other structures) and provides auto schematization for quick exploration of the data.

In ingesting data, users first set up a Flex Table with a command such as:

```
CREATE FLEX TABLE tweets ();
```

Then users can parse the data with a single command such as:

```
COPY tweets FROM '/path/file.gz' GZIP PARSER fjsonparser();
```

As of Version 7, two parsers support loading Flex Tables, fdelimitedparser and fjsonparser. Both parsers store the data as a single value map. You can use parsers in succession to load data. For example, you could load data from a delimited file and then from JSON in succession into a single value map.

Having a function like Flex Tables that automatically understands common schemas is powerful. It takes time to map structure and set up schemas, slowing the overall process of exploration of the data. Should the structure of the data change, maintaining the schemas is also time-consuming. By integrating these less structured data sources and supporting vanilla SQL queries against them, we bring a key feature of relational databases to bear: abstracting the storage representation from the query semantics.

Once you've explored your dark data, there may be some data you want to materialize into columnar tables to make it part of your regular analytics. You can materialize tables in Vertica in one step for an additional increase in analytical performance by taking advantage of Vertica's compression and columnar storage architecture.

Flex Tables Storage

While loading data into Flex Tables, data is stored in a single column and this effectively turns Vertica into a row store. This offers subtle differences over native columnar store:

- Data is easily ingested and excellent for exploring unknown data. The schema is automatic, based on the information provided in the raw data. You can spend less time dealing with the ETL process and changing dimensions.
- •While the data is accessible using similar methods to the native storage, the basic storage model lacks many advantages of the data exploration performance of Vertica. Data is not maximized for I/O utilization, optimal disk footprint is absent, and it does not provide indexing functionality.

VIRTUAL SCHEMAS AND VIEWS

After loading data into a Flex Table, one of the first tasks most users perform is to understand what key-value pairs exist as populated virtual columns in the data. The `COMPUTE_FLEXTABLE_KEYST` function computes keys from map data. You can also compute the keys and generate a view with the `COMPUTE_FLEXTABLE_KEYS_AND_BUILD_VIEW`.

With Vertica Flex Tables, the Vertica Analytics Platform now enables you to perform data exploration on semi-structured data. Building on the capabilities of Vertica, including its extremely fast and scalable database, Flex Tables offer blazingly fast analytics on semi-structured and structured data. It offers more complete, better performing SQL queries than SQL on Hadoop.

SQL ON FLEX TABLES

Once data is available in Flex Tables, you can easily tap into the power of Vertica and query your data. It offers the same fast and powerful SQL to perform data analytics and queries on Flex Tables as easily and transparently as you query your Vertica datastore.

Vertica provides the richest most open SQL on Hadoop. You can sometimes encounter major limitations in certain SQL and NoSQL solutions. For example, Vertica supports JOIN and the use of the set operators—INTERSECT, EXCEPT, and UNION—are sometimes missing or limited on Big Data analytical databases. So, if you need to combine rows from two or more tables based on a common field, you may be prevented from achieving this.

We encourage the use of your visualization tools with your Vertica Flex Tables. The combination of data visualization and Flex Tables delivers insights quickly to explore your dark data. A list of partners for data visualization and business intelligence can be found on our website on the business intelligence partners page.

CREATING COLUMNAR TABLES FROM FLEX TABLES

Once you've explored your dark data, there may be some data you want to materialize into columnar tables to make it part of your regular analytics. You can materialize tables in Vertica in one step for an additional increase in analytical performance by taking advantage of Vertica's compression and columnar storage architecture. Vertica makes it easy to create a regular columnar table from a Flex Table. Typically, you specify the virtual columns you want to materialize into a regular table. For example, if the Flex Table has two virtual columns (user.lang and user.name) that you want to materialize in a columnar table, enter a command such as the following:

```
create table darkdata_full as select "user.lang"::varchar,
"user.name"::varchar from darkdata;
```

External Tables

Vertica also supports external tables where the data resides outside the database and is loaded at query time. You can use the Hadoop Distributed File System (HDFS) Connector as a source for an external table that lets you directly perform data analysis queries on the HDFS contents. This feature is best implemented for data that's needed for infrequently run queries. If you need to query the content of the HDFS files frequently, it's better to load the entire content of the files into Vertica. If your HDFS data is in the optimized row columnar (ORC) format and uses no complex data types, you can use the ORC Reader to access the data directly. Reading directly may provide better performance.

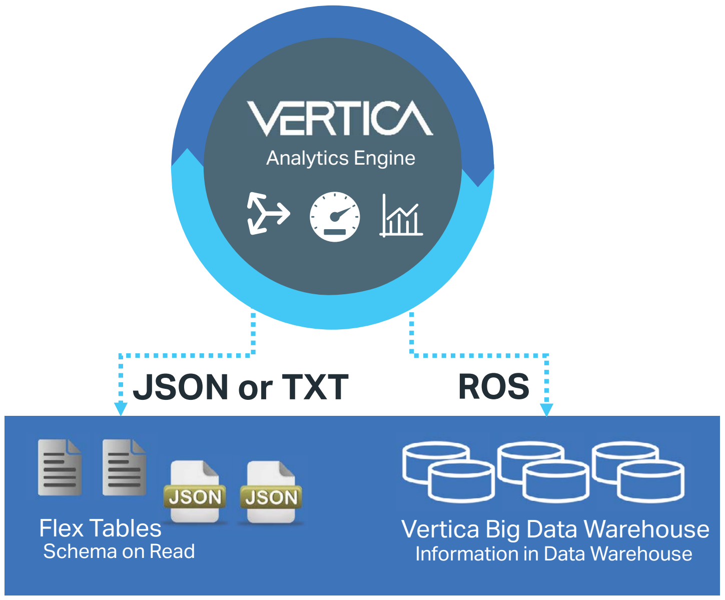

Figure 1. Vertica Analytics Engine

Real-World Use Cases

Get quick start and full documentation for real-world examples at the My Vertica Community website.

Boosting Performance

It's incredibly easy to go from dark data exploration in Flex Tables to daily analytics in native Vertica. When you're ready to use the data as part of your daily analytics, a simple promotion from exploration to production will let you analyze your data fast. Use Flex Tables to discover the value of your semi-structured data. When you find it, you can materialize tables with one step for an additional increase in analytical performance by taking advantage of Vertica's compression and columnar storage architecture.

While Flex Tables are great for exploration, you will probably want to do some tuning before moving into production. Luckily, tuning is easy and still preserves the flexible nature of the tables. The central mechanism is "promotion" of a column to being a real column of the table. Typically, you select commonly used columns, particularly those used in joins or predicates. You can do this by using `ALTER TABLE ... ADD COLUMN ... DEFAULT`. For example, the time and vote columns might be of particular interest:

```
ALTER TABLE webreqs ADD COLUMN "time" TIMESTAMP NOT NULL DEFAULT
time::timestamp;

ALTER TABLE webreqs ADD COLUMN "data.vote.0" VARCHAR(10) DEFAULT
"data. vote.0"::varchar;
```

As part of adding the column, Vertica populates the column with data. Future loads will also load the newly added columns. You will note that we have selected specific data types for these columns. New data that doesn't match the declared data type will result in nulls getting stored (or errors if the column is declared "NOT NULL").

Vertica's Database Designer helps you to optimize the physical design (storage layout) and it works great on promoted columns in Flex Tables. A typical usage model is explore, promote, and optimize. A key thing to remember is that changes to the schema do not require promotion and optimization. If you add fields that dramatically change your query workload, then some additional performance optimization may be necessary.

Summary

With Vertica Flex Tables, the Vertica Analytics Platform now enables you to perform data exploration on semi-structured data. Building on the capabilities of Vertica, including its extremely fast and scalable database, Flex Tables offer blazingly fast analytics on semi-structured and structured data. It offers more complete, better performing SQL queries than SQL on Hadoop.

Flex Tables emphasize ease of use, particularly around managing data schemas. Auto-schematization simplifies your process of loading and exploring data and frees you from having to spend technical resources on understanding and managing schemas. It removes the need for coding-intensive schemas to be defined or applied before the data is loaded for exploration. It also offers "one-click" schema capabilities that allow schemas to be created and applied as needed for high-performance analytics. This enables analysts, data scientists, and business users to explore and visualize information without burdening, or needing to wait for, their IT organizations to use the lengthy and expensive extraction and transformation tools and processes typical with legacy databases and data warehouses.

Learn More At
https://www.vertica.com/

Additional contact information and office locations:
www.vertica.com

www.vertica.com