An Indexing Approach to Chemical Data Management

Indexing is a well-understood and broadly applied approach to data integration. Google search is probably the best example. Independent of any underlying data structure, indexing services like Google "ingest" text and make the contents available for incredibly fast search, with links to the underlying source systems. Scoring systems and relevancy metrics make indexing an ever more powerful tool to get most relevant search results within seconds.The use of modern, no SQL type technologies has been largely driven by the desire for a key characteristic, namely the implementation of horizontally scalable indexing. By virtue of this capability to spread search and retrieval across essentially an infinite number of worker nodes, it has been possible to develop systems of enormous scale – scale never before even contemplated. This is an essential capability for cloud- scale informatics tools.

Licensing models for most indexing systems are also attractive and aimed at cloud-scale computing. Many indexing systems are open source, and even in cases where they are licensed, license fees for cloud deployments across hundreds of compute nodes tend to be much cheaper and more flexible than their relational database analogs. While this is a guideline more than a rule, it is certainly true for Oracle platforms.

However, indexing technologies run afoul of certain fundamental challenges. Take this search task for example:

> "Find all X where X is joined to Y and attributes of X are in a certain range and attributes of Y are in certain range".

While a relational database would (generally) be able to return that result quickly and precisely, indexing technologies often cannot. Even in this simple example, most indexing systems underperform or even fail, because they either can't handle the quantitative precision of the attribute search or the 'joining' is cumbersome to implement and substantially mitigates the benefits of the entire approach in terms of both scalability and effort.

Another fundamental challenge has to do with extensibility of search. In mature relational database platforms, API's exist to embed new object classes within the RDBMS engine. 3D spatial searching is the best out-of-the-box example of object extensibility in both the Oracle and Microsoft SQL*Server platforms. In our industry, the best examples are the capability to embed chemical structure and sequence searching within the Oracle RDBMS system using the Oracle cartridge API.

Until now, this kind of object extensibility hasn't been available within indexing systems to the same degree. As companies seek to transition to Cloud computing, the gap becomes more important as it is typically not cost effective or even technically possible to transition these scientific search capabilities directly into the Cloud.

Bringing Index Based Search to Scientific

Computing Applications Revvity Signals™ Data Factory represents a new kind of solution that brings the benefits of index-based search to scientific computing applications while at the same time maintaining the benefits of hyper-scalability and ease of implementation.

There are three highly-innovative aspects to Revvity Signals Data Factory that make this possible:

1. Capture search constrains and search attributes: Revvity Informatics has developed a patent-pending algorithm to enable chemical structure search within Apache Lucene-based indexing systems (Lucene is the underlying technology for modern indexing frameworks such as Elastic Search). This means that within a single query, it is possible to capture chemical-structure search constraints along with other search attributes. The net result is extraordinarily fast (near-real-time) structure search ideally suited for Cloud computing applications.

2. Suitable for Chemical Property and Test Data: Revvity Signals Data Factory is designed to enable scientists to discovery the correlation between variations of composition or material morphology and their resulting properties. In addition to its unique structure search capabilities, Revvity Signals Data Factory includes capabilities to shape and annotate formulation and materials testing data into a hyper-scalable index that enables precise quantitative search seamlessly integrated with structure search. For example, it is possible to execute queries like the following with extreme speed: "retrieve all test results for formulations containing a certain chemical structure where the observed physical properties in one or more tests lie within a desired range."

3. No costly ETL skills needed; The architecture of the underlying indexing technology is based on nimble and light-weight information design tools that bypass rigid schema definition and costly coded Extract-Transform Load approaches typically used in traditional data warehouse projects. Programming teams can gain access to data from any business intelligence or computational tool that can issue REST API requests. These factors taken together make Revvity Signals Data Factory the first composition-activity correlation application to provide the full-benefits of cloud-scale index-based search to scientific computing applications.

Additional Background

Revvity Signals is committed to the notion that technical partnerships are essential in this era as it is impossible to develop all required capabilities on one's own. As such, we are committed to both commercial strategic partnerships and the inclusion of open source technology throughout our platform. Revvity Signals Data Factory is built on top of open source technologies such as Elastic Search and Spark, while fully integrated with our partner's technology, such as TIBCO Spotfire®, the premier visual analytics platform for science.

The primary graphical interface of Revvity Signals Data Factory is Spotfire®. Revvity Signals Data Factory fills the lack of search back-end in Spotfire® by marrying cloud-scale search to Spotfire's inherent strength: allowing scientists to find insights, quickly. Spotfire®customers will be able to easily integrate Revvity Signals Data Factory into their IT landscape using the systems data ingestion tools and automation REST API's. Revvity Signals Data Factory provides an extraordinarily rich data harmonization back-end. It uses a hyper-scalable Spark cluster for data staging, aggregation and

AI/ML workloads. It includes an Elastic Search cluster with embedded scientific intelligence that enables scientists to seamlessly find and retrieve large and complex datasets for detailed analysis and computation. While Spark and Elastic Search open source projects are widely available and used by data science groups, Revvity Signals has integrated and tunned these general purpose technologies into a truly innovative out-of-the-box solution for scientific data analysis. The key value of the solution lies in the innovative information design, meta-data capture, ingestion, and mapping tools developed by Revvity Signals to address the specific challenges of experimental data management.

In many respects "standing on the shoulders of giants", Revvity Signals Data Factory is a a modern application for today's new science.

revvitysignals.com
940 Winter Street
Waltham, MA 02451 USA
P: (800) 762-4000 (+1) 203-925-4602

Revvity Signals
RevvitySignalsSoftware
revvitysignals
Revvity_Signals
RevvitySignals