

РАЗРАБОТКА СИСТЕМЫ ИНТЕГРАЦИИ МЕТАДААННЫХ НАУЧНОГО ЭКСПЕРИМЕНТА

М. Ю. Губин, М. А. Григорьева, М. В. Голосова,
Томский политехнический университет, НИЦ «Курчатовский институт»
mgubin@tpu.ru

Введение

Современные научные эксперименты часто являются длительными, географически распределёнными, требуют обработки и хранения большого количества собранных научных данных (порядка петабайт и эксабайт) в течение всего времени проведения эксперимента.

Порождаемые на всех стадиях жизненного цикла эксперимента данные сопровождаются метаданными, необходимыми для отслеживания хода эксперимента, обеспечения целостности экспериментальных данных, а также для обеспечения воспроизводимости эксперимента в будущем.

В большинстве научных сообществ метаданные, описывающие цепочки анализа и обработки данных, и метаданные о публикации научных результатов создаются и хранятся независимо друг от друга, часто - в различных хранилищах. Кроме того, чтобы воспроизвести или подтвердить результаты уже проведенного эксперимента, ученым бывает необходимо провести исследования при тех же условиях, проверить результаты обработки наборов данных новой версией программного обеспечения, или опробовать новые алгоритмы. Это приводит к необходимости длительно хранить и иметь возможность анализировать большое количество разнородных метаданных, начиная от выдвигаемой гипотезы и цепочки преобразования данных, и до публикации результатов.

Целью нашей работы является разработка системы интеграции метаданных научного эксперимента, которая призвана предоставить через единый интерфейс доступ ко всем метаданным, которые были порождены за все время проведения эксперимента, и обеспечить поиск и анализ метаданных различных типов и полученных из разнородных источников в пределах одного поискового запроса.

Описание системы интеграции метаданных

Решение проблемы интеграции разнородных метаданных из различных источников требует использования хранилища, рассчитанного на работу с сильно связанными данными со сложной структурой, часто подвергающейся изменениям.

В качестве основы такого хранилища была выбрана онтологическая база данных OpenLink Virtuoso, поскольку в тестах производительности и масштабируемости на объемах данных, сравнимых с предполагаемыми объемами метаданных научного эксперимента, оно показало хорошие результаты. По данным сравнительного анализа,

проведенного с помощью Berlin SPARQL Benchmark (BSBM), Virtuoso выигрывает у своих ближайших конкурентов (BigData, BigOWLIM и TDB) по количеству различных запросов от одного и от многих (до 64-х) клиентов, выполненных за 1 час, а также по количеству запросов в секунду, выполненных на различных объемах данных (100 млн., 200 млн., 1 млрд. триплетов). [5]

Архитектура разработанного хранилища метаданных представлена на рисунке 1.

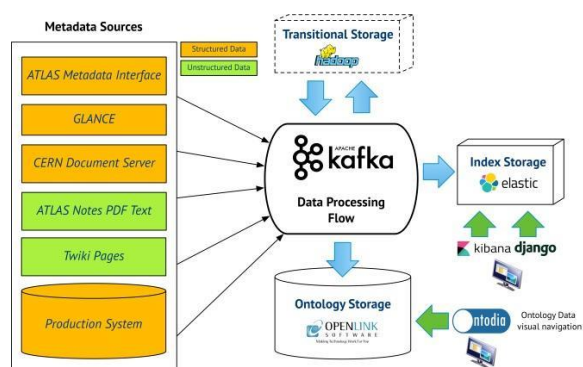


Рис.1. Архитектура хранилища метаданных

Система собирает метаданные из множества разнородных источников (CERN Document Server, InSpire, страниц Twiki и JIRA, системы планирования совещаний Indico, баз данных JEDI, DEFT и Rucio) с применением системы потоковой обработки данных Kafka. Данные, требующие вычислительно тяжелой предварительной обработки, помещаются в промежуточное распределённое хранилище на базе стека технологий Apache Hadoop, где производится их подготовка и преобразование в формат, подходящий для загрузки в центральное онтологическое хранилище Virtuoso. Но основная часть данных обрабатывается в потоковом режиме и загружается в Virtuoso, минуя промежуточное хранилище, что позволяет сократить время обработки. Поточковая обработка данных управляется средствами Apache Kafka. Данные, к которым требуется быстрый и частый доступ с целью их простого анализа, индексируются в системе полнотекстового поиска Elasticsearch. Для визуализации данных на промежуточном этапе их обработки используется Kibana, а для представления их конечному пользователю применяется Ontodia – специализированный инструмент для навигации по онтологическим данным. [3]

Онтология эксперимента ATLAS

В качестве механизма интеграции разнородных данных в единую связанную систему был выбран набор подходов Semantic Web. Для каждого эксперимента, который будет использовать хранилище, необходима разработка специализированной онтологии предметной области, которая описывает связи между информацией из разнородных источников, и позволяет производить автоматизированную интеграцию поступающей информации в единое связанное целое.

Разработанная онтологическая модель описывает структуру связей между важными для описания хода научного эксперимента элементами предметной области, такими как научные публикации, черновики публикаций, журналы, данные об авторах документов, идентификаторы и описания научных данных, которые легли в основу публикации (такие, как имя набора данных, название проекта физического анализа и кампании обработки данных, название физической группы, выполнившей анализ, а также некоторые параметры оборудования, такие как тип столкновения и энергия столкновения). [1, 3] В дальнейшем набор общих атрибутов будет увеличиваться, что позволит обеспечить более сильную связность между метаданными.

Существующая онтология постоянно развивается и дорабатывается, чтобы отражать современное состояние предметной области и позволять осуществлять анализ данных, необходимый ученым, проводящим эксперимент, а также аналитикам.

Адаптация существующей модели к новым экспериментам

Опыт разработки онтологической модели эксперимента ATLAS показал, что создание онтологий предметных областей является длительным процессом, требующим значительных трудозатрат.

Для снижения трудозатрат при разработке онтологической модели были рассмотрены подходы для автоматического построения моделей предметных областей. [6]

Модели предметной области, которые можно строить автоматически, часто генерируются с ошибками, которые в дальнейшем могут приводить к выдаче неправильных данных на запросы пользователей. Хотя данное направление является перспективным, на данном этапе его развития оно не дает выигрыша в разработке моделей предметных областей, для которых необходима высокая точность, удобство для пользователя, и отсутствие ошибок в структуре модели.

Заключение

На данном этапе своего развития, DKB представляет собой прототип системы интеграции метаданных научного эксперимента, и уже реализованный функционал позволяет судить о перспективности исследуемого подхода к хранению метаданных. Можно ожидать, что полностью реализованное хранилище значительно упростит воспроизведение экспериментов, метаданные о которых были внесены в него, а также облегчит установление связи между публикациями в научных журналах и экспериментальными данными, на основе которых были выпущены эти публикации. Также стали видны и некоторые недостатки подхода, в частности необходимость разрабатывать подробную онтологию для каждой новой предметной области существенно замедляет реализацию хранилища.

Дальнейшее развитие DKB будет связано с расширением онтологической модели, подключением новых источников метаданных, разработкой сервисов и рабочих потоков обработки данных, тестированием подходов автоматической генерации моделей предметных областей, и усовершенствованием механизмов анализа PDF документов.

Список использованных источников

1. M. A. Grigorieva, M. V. Golosova, M. Y. Gubin, A. A. Klimentov, V. V. Osipova, E. A. Ryabinkin Evaluating non-relational storage technology for HEP metadata and meta-data catalog [Electronic resource] // Journal of Physics: Conference Series. — 2016. — Vol. 762, iss. 1 : Advanced Computing and Analysis Techniques in Physics Research (ACAT2016). — [012017, 5 p.].
2. Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Elsevier. — 2011. ISBN 978-0-12-385965-5.
3. Григорьева М., Голосова М., Рябинкин Е., Климентов А. Экзабайтное хранилище научных данных // Открытые системы. СУБД. — 2015. — Т.4. — С. 14-17
4. Erling O., Mikhailov I. RDF Support in the Virtuoso DBMS //Networked Knowledge-Networked Media. — Springer Berlin Heidelberg, 2009. — С. 7-24.
5. Berlin SPARQL Benchmark results [Электронный ресурс] / Режим доступа: <http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/index.html> свободный.
6. Balakrishna M., Moldovan D. I. Automatic Building of Semantically Rich Domain Models from Unstructured Data //FLAIRS Conference. — 2013.