Научная визуализация, 2018, том 10, номер 1, страницы 77 - 88, DOI: 10.26583/sv.10.1.06

Новый подход к мониторингу системы управления потоками заданий ProdSys2/PanDA эксперимента ATLAS на Большом адронном коллайдере с использованием средств и методов визуальной аналитики

Т.П. Галкин[A,1], М.А. Григорьева[B,D,2], А.А. Климентов[B,C,3], Т.А. Корчуганова[D,4], И.Е. Мильман[A,5], С.В. Падольский[C,6], В.В. Пилюгин[A,7], Д.Д. Попов[A,8], М.А. Титов[B,9]

[A] Национальный исследовательский ядерный университет “МИФИ”, Россия
[B] Национальный исследовательский центр “Курчатовский институт”, Россия
[C] Брукхейвенская Национальная Лаборатория, США
[D] Национальный исследовательский Томский политехнический университет, Россия

[1] ORCID: 0000-0003-2859-6275, TPGalkin@mephi.ru
[2] ORCID: 0000-0002-8851-2187, magsend@gmail.com
[3] ORCID: 0000-0003-2748-4829, Alexei.Klimentov@cern.ch
[4] ORCID: 0000-0001-5792-8182, tatiana.korchuganova@cern.ch
[5] ORCID: 0000-0001-9705-9401, igalush@gmail.com
[6] ORCID: 0000-0002-6795-7670, spadolski@bnl.gov
[7] ORCID: 0000-0001-8648-1690, VVPilyugin@mephi.ru
[8] ORCID: 0000-0002-3333-749X, DDPopov@mephi.ru
[9] ORCID: 0000-0003-2357-7382, mikhail.titov@cern.ch

Аннотация

В статье представлен пилотный проект "Средства и методы визуальной аналитики как часть системы управления потоками заданий и загрузкой ProdSys2/PanDA эксперимента ATLAS на Большом адронном коллайдере". Проект направлен на расширение функциональных возможностей существующей системы мониторинга эксперимента ATLAS используя подход визуальной аналитики для анализа больших объемов многомерных данных о вычислительных задачах. Функционирование систем распределенной обработки и анализа данных эксперимента ATLAS связано с обработкой мультипетабайтных и эксабайтных объемов данных. При этом возникают задачи, требующие применения средств многоуровневой интерактивной визуализации для анализа корреляций между отдельными наборами данных и их представлениями. Статья содержит описание предлагаемых в проекте подходов к визуальному анализу многомерных данных, а также определение областей применения визуальной аналитики при обработке данных в эксперименте ATLAS. Определены задачи для апробации предложенных методов на примере анализа статистических данных системы управления загрузкой эксперимента ATLAS. Заключительная часть статьи посвящена организационной составляющей пилотного проекта.

Ключевые слова: визуальная аналитика, система управления потоками заданий, большие данные, БАК

1. Введение

Описываемый в статье пилотный проект “Средства и методы визуальной аналитики как часть системы управления потоками заданий и загрузкой ProdSys2/PanDA эксперимента ATLAS на Большом адронном коллайдере” направлен на разработку современных подходов и инструментов визуализации и аналитики для мониторинга и контроля системы управления распределенными потоками заданий и загрузкой (англ. Workflow Management System) на примере системы обработки и анализа

данных эксперимента ATLAS [1] на LHC [2] (Большой адронный коллайдер, ЦЕРН, Швейцария). Начальная мотивация проекта связана с экспериментами на LHC, но как количественные, так и качественные требования являются общими для многих экспериментов в области физики высоких энергий (ФВЭ) и ядерной физики (ЯФ), поэтому данный проект интересен широкому кругу научных групп, как показало обсуждение в рамках международной конференции Nuclear Electronics and Computing (NEC2017) [3].

Внутренняя информация (такая как, описание формата и процессов обработки данных физических экспериментов) распределенных систем обработки данных также должна обрабатываться, анализироваться и представляться пользователям системы в компактной форме. Для этого разрабатываются специализированные системы контроля и мониторинга. В статье [4], также опубликованной в этом номере журнала, описывается существующая в эксперименте ATLAS система мониторинга - BigPanDA. Система включает в себя следующие возможности для визуализации данных: интерактивные интерфейсы, параметрические таблицы и различные форматы графиков (гистограммы, линейчатые и круговые диаграммы, простые двумерные графики). До настоящего времени требования, предъявляемые к системе мониторинга, ограничивались использованием базового визуального анализа для заданных классов задач, и данных, размерностью не более трех измерений. Однако, постоянное увеличение объемов обрабатываемых данных и усложнение вычислительной инфраструктуры обработки данных эксперимента ATLAS, как будет показано ниже, порождает новые задачи, связанные с визуальной аналитикой больших объемов многомерных данных.

В рамках данного проекта предлагается применить принципиально новый подход к контролю работы сложных распределенных систем и изменить “классический” мониторинг систем обработки данных (в частности, в области ФВЭ и ЯФ), используя методы визуальной аналитики [5,6] для повышения качества оценочных показателей состояния данных. Применение данных методов позволит снять ограничение по размерности анализируемых данных, обеспечив построение многомерных геометрических интерпретаций для визуального анализа. Как результат, функциональность системы мониторинга будет расширена возможностями выявления явно выраженных корреляций между различными объектами данных и использованием многоуровневых интерактивных интерфейсов. Кроме того, совместное применение методов визуальной аналитики и “машинного обучения”, позволит повысить уровень автоматизации при эксплуатации систем обработки и анализа данных в первую очередь в области физики элементарных частиц.

2. Проблемы контроля и управления сверхбольшими данными

Цикл научных исследований на современных установках составляет десятилетия (например, научные сообщества на LHC были созданы более 25 лет назад, набор данных был начат 9 лет назад и будет продолжаться по крайней мере еще 10 лет). За время жизни экспериментов происходит модернизация установок и детекторов, качественно меняются информационные технологии. Одновременно меняется и программно-аппаратная инфраструктура и модель обработки данных (компьютерная модель): постоянно увеличивается количество вычислительных центров (появляются новые архитектурные решения), меняются сценарии запуска и выполнения задач анализа, обработки и моделирования данных, обновляются версии программного обеспечения, меняются технологии хранения данных и доступа к данным и метаданным. В условиях постоянно эволюционирующей и усложняющейся вычислительной инфраструктуры и одновременном росте потока информации становится все

сложнее контролировать работу систем управления данными, планировать обработку данных и моделирование эксперимента, прогнозировать и своевременно обнаруживать возможные аномалии в работе отдельных аппаратных и программных компонент вычислительной инфраструктуры, и систем для обработки данных в целом. Для решения этих проблем в современных экспериментах разрабатываются специализированные аналитические инструменты. Отдельной сложной задачей является поиск закономерностей или анализ аномалий в работе комплексных распределенных систем, корреляций между действиями операторской службы и поведением системы, предсказание поведения системы в случае изменений программного обеспечения.

Второе поколение системы обработки, моделирования и анализа данных ProdSys2 (англ. Production System) [7] эксперимента ATLAS совместно с системой управления загрузкой PanDA [8] (англ. Production and Distributed Analysis system) - это сложный комплекс аппаратно-программных средств для организации, планирования, запуска и выполнения вычислительных задач (Рис. 1). ProdSys2/PanDA отвечает за все этапы обработки, анализа и моделирования данных, в том числе моделирование физических процессов и работы детектора методом Монте-Карло, (пере)обработку физических данных, выполнение узкоспециализированных заданий (например, задания триггера высшего уровня или задачи проверки качества ПО). Используя ПО ProdSys2/PanDA научное сообщество ATLAS, отдельные физические группы и ученые имеют доступ к сотням вычислительных центров консорциума WLCG [9] (англ. Worldwide LHC Computing Grid), суперкомпьютерам, ресурсам облачных вычислений и университетским кластерам. Характеристики системы могут быть отражены следующими показателями: выполнение более миллиона задач в день на 200+ вычислительных центрах (ВЦ) тысячью пользователями с использованием более 300 тысяч узлов.

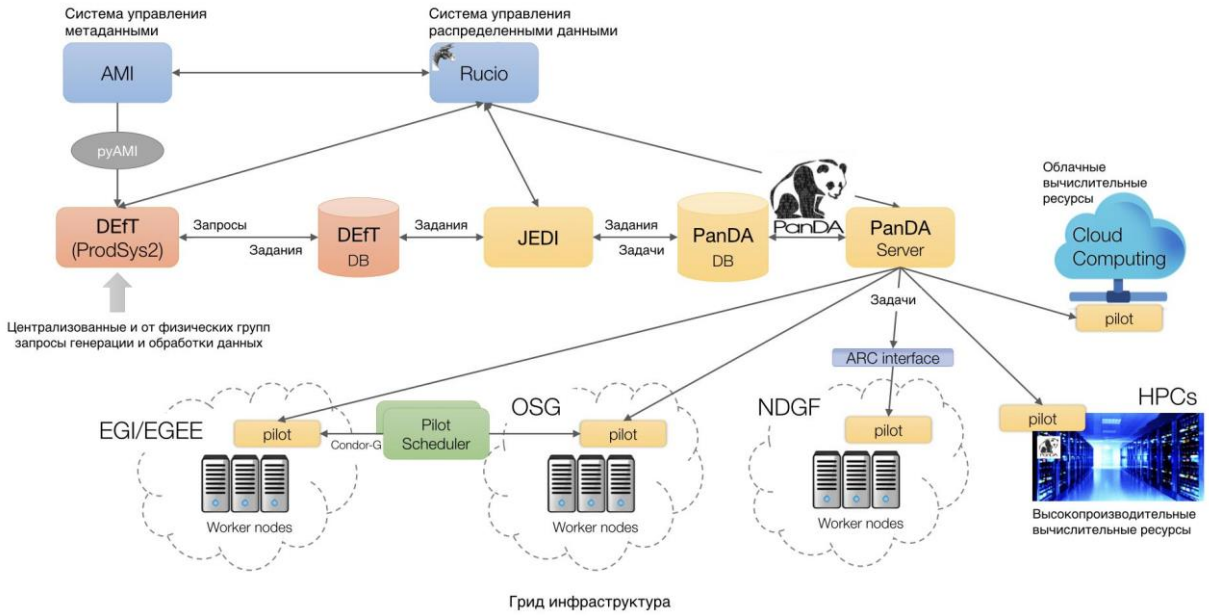


Рис.1. Рабочий поток обработки данных в эксперименте ATLAS

В распределенной системе постоянно существует конкуренция между различными потоками заданий. Например в период, предшествующий основным физическим конференциям резко возрастает количество задач анализа данных (из рисунка 2 следует, что в эксперименте ATLAS происходило резкое

увеличение числа выполняемых задач в определенные месяцы). Поэтому при запуске задач могут возникать существенные задержки, связанные с отсутствием свободного вычислительного ресурса. Как правило, конечного пользователя интересует не столько процесс выполнения задач, сколько возможность предсказать время окончания обработки (или анализа) данных и получить научный результат в заранее определенный срок (это могут быть часы для задач анализа, или недели при переобработке данных). Сам поток выполнения заданий имеет несколько фаз выполнения (например: моделирование, оцифровка, реконструкция, создание объектов для физического анализа, физический анализ) и каждый этап (фаза) может выполняться на географически разделенных ВЦ, что включает в себя передачу исходных данных между ВЦ и может влиять на общее время выполнения всего потока заданий. Возможные аппаратные сбои могут потребовать перераспределения заданий между ВЦ, что также создает дополнительную неоднозначность в предсказании процесса обработки данных. В настоящий момент не существует центрального портала для контроля, оператор вынужден просматривать графики передачи данных, графики выполнения заданий, таблицы с информацией о работе ВЦ и отдельных компонент. Создание единого портала и возможность визуализации работы систем позволит оптимизировать использование вычислительного ресурса, значительно автоматизировать и упростить процесс обработки данных, и тем самым ускорить получение научного результата.

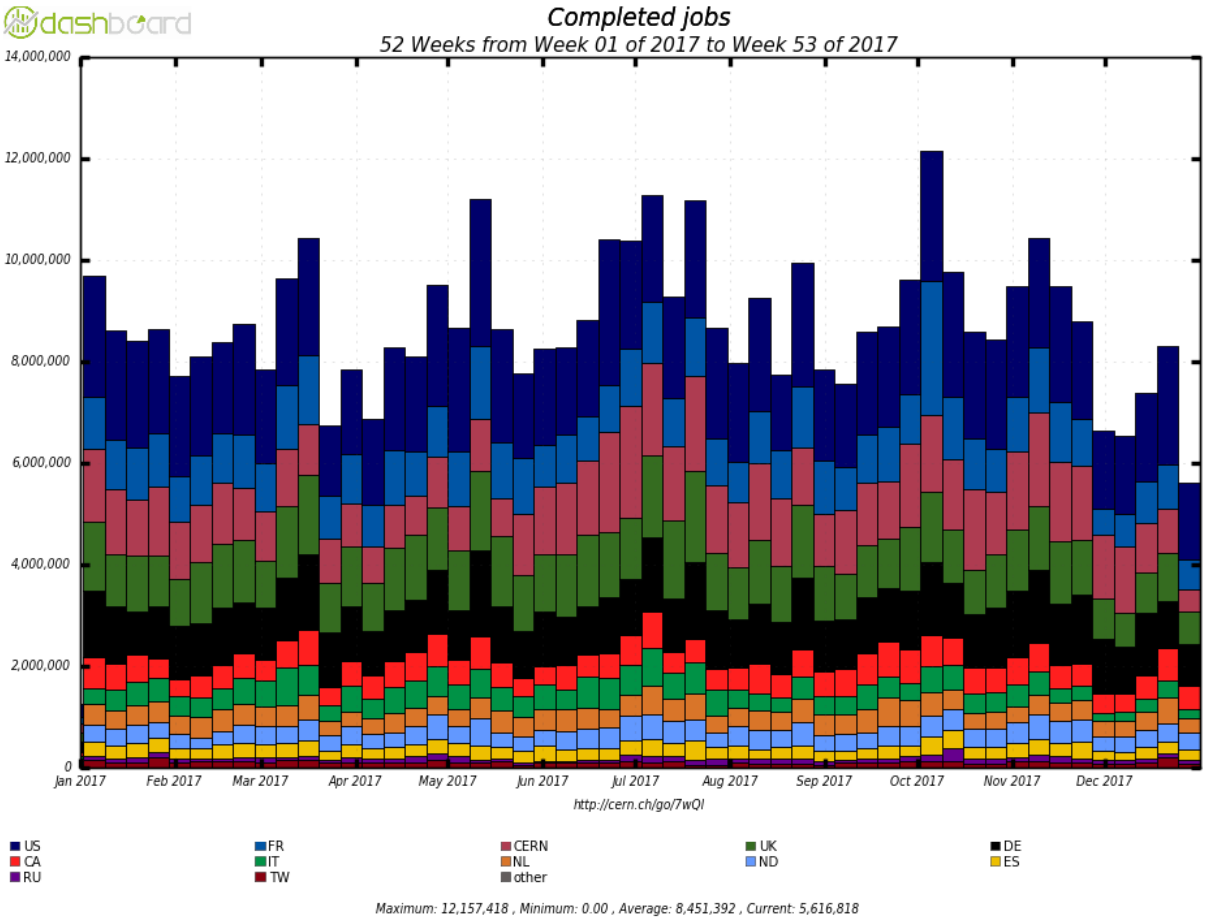


Рис.2. Количество выполненных задач по группам вычислительных центров в течение 2017 г. (среднее за неделю)

Информация накопленная за весь период функционирования системы управления заданиями и загрузкой в эксперименте ATLAS (а это более 14 лет), содержит данные о ходе выполнения более чем 10 миллионов заданий и

около 3000 миллионов задач. Такая статистика позволяет использовать методы “машинного обучения” для аналитических расчетов и прогнозирования работы программного обеспечения. Работы, связанные с ProdSys2/PanDA, в которых были начаты разработка и применение методов "машинного обучения" для анализа данных обработки [10,11], показали, что рассчитанные метрики на основе прогнозируемой аналитики позволяют повысить эффективность процессов обработки физических и моделируемых данных - более тщательное планирование процесса анализа (определяется индивидуальными пользователями или группой пользователей), прогнозирование возможного отказа или аномального поведения системы (производится агентами служб контроля).

3. Подход визуальной аналитики

В современном мире проблемы обработки и анализа многомерных данных являются одними из актуальнейших задач. Для решения подобных задач разрабатывается множество различных методов и программно-аппаратных средств, как автоматических, так и интерактивных. Следует отметить, что в рамках этих методов может использоваться визуализация данных. В настоящее время широко используется подход визуальной аналитики. Этому подходу предшествовали методы решения задач анализа многомерных данных различными визуальными методами.

Изучение литературы, посвященной описанию конкретных приложений с применением визуальных методов, позволяет утверждать, что в реальности интерактивным системам работы с многомерными данными зачастую придается меньшее значение по сравнению с системами отображения результатов применения методов анализа данных (англ. data analysis). В качестве примера можно привести такие системы, как система ситуационного оповещения AdAware [12], система визуального анализа в задачах самолетостроения [13], программный комплекс SAS Visual Analytics [14], предназначенный для обработки и анализа больших объемов финансовой и экономической информации. Как показывает практика, для визуального представления многомерных данных широко используются классические методы параллельных координат, кривых Эндрюса, лиц Чернова и других подобных мнемонических графических отображений. Все эти визуальные методы основаны на том, что анализируемые кортежи числовых данных интерпретируются в качестве значений параметров таких мнемонических графических отображений. Примеры таких отображений представлены на рисунках 3,4.

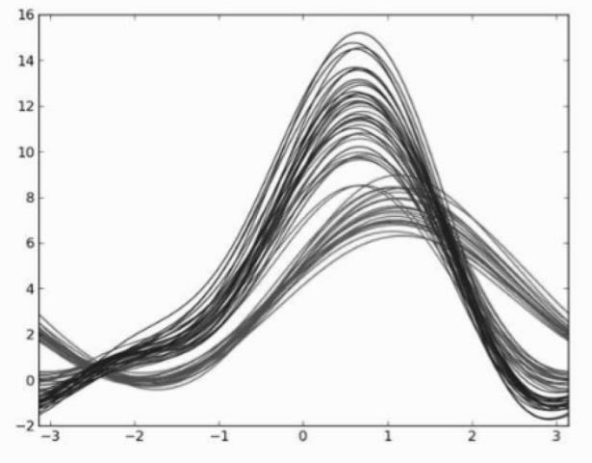


Рис.3. Ирисы Фишера, представленные в виде кривых Эндрюса [15]

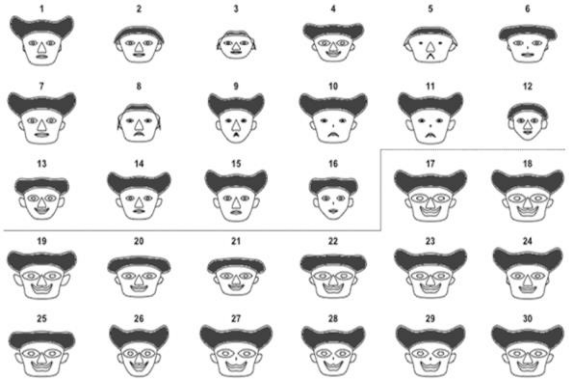


Рис.4. Лица Чернова для медицинских данных [16]

Следует отметить, что все системы, использующие упомянутые визуальные методы, по сути настроены на внутреннюю обработку многомерных данных и представление их аналитику в удобном

для него виде. Они не предоставляют возможности непосредственно работать с анализируемыми данными и их, естественными для человека, многомерными геометрическими интерпретациями с использованием визуальных отображений этих интерпретаций. Подход визуальной аналитики предполагает решение задач анализа данных, и в частности задач анализа многомерных данных, с использованием способствующего интерактивного визуального интерфейса.

Одной из наиболее распространённых форм визуальной аналитики является решение задач анализа многомерных данных методом визуализации [17]. Решение задачи анализа исходных данных методом визуализации заключается в последовательном решении 2-х следующих задач (Рис.5).

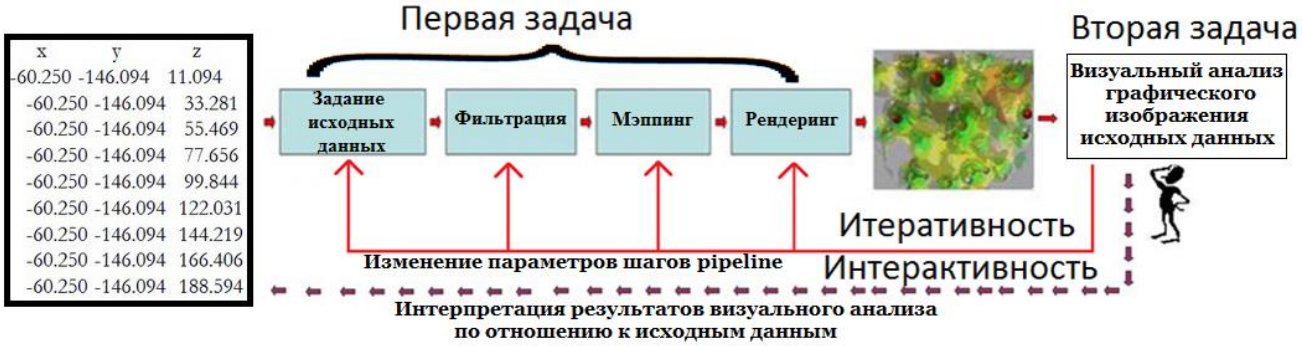


Рис.5. Анализ данных методом визуализации

Первая задача заключается в получении представления анализируемых данных в виде их некоторого графического изображения (задача визуализации исходных данных). Эта задача решается с использованием компьютера. Получаемые графические изображения служат естественным и удобным средством представления пространственной интерпретации исходных данных человеку (аналитику). Пространственная интерпретация представляет собой один или несколько пространственных объектов (так называемая, пространственная сцена), которые ставятся в соответствие анализируемым данным. Вторая задача, которая является не менее важной, заключается в визуальном анализе графического изображения анализируемых данных, полученного в результате решения первой задачи, при этом результаты анализа интерпретируются по отношению к исходным данным. Эта задача решается непосредственно аналитиком. Пространственная сцена визуально анализируется, используя огромные потенциальные возможности пространственно-образного мышления аналитика в процессе анализа. В результате решения данной задачи аналитик делает некоторые суждения о пространственной сцене. Таким образом, формулируются суждения о рассматриваемом объекте. Процесс визуального анализа графического изображения строго не формализуем. Эффективность визуального анализа определяются опытом человека, осуществляющего этот анализ изображения, и его склонностью к пространственно-образному мышлению. Глядя на полученное изображение, человек может решать 3 основные задачи: анализ формы пространственных объектов, анализ их взаимного расположения и анализ графических атрибутов пространственных объектов.

В рамках данного проекта предлагается использование многомерного геометрического моделирования исходных данных, которые рассматриваются как многомерные табличные данные о вычислительных задачах (Табл. 1).

Табл.1. Данные о вычислительных задачах

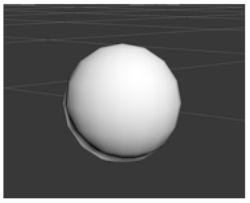
	Параметр 1	Параметр 2	…	Параметр n
Задача 1	x_1^1	x_1^2	…	x_1^n
…	…	…	…	
Задача i	x_i^1	x_i^2	…	x_i^n
…	…	…	…	…
Задача m	x_m^1	x_m^2	…	x_m^n

Для решения поставленной задачи проводится геометрическая интерпретация. Строкам таблицы ставятся в соответствие многомерные точки в пространстве En, $p_i = (p_i^1, p_i^2, \ldots, p_i^n) \in E_n$, а значения параметров задач - это координаты многомерных точек. Меру различия строк параметров предлагается интерпретировать как евклидово расстояние между точками этого многомерного пространства (чем больше расстояние, тем больше различаются строки). При такой интерпретации, задаче анализа схожести и различия вычислительных задач ставится в соответствие задача анализа расстояния между точками n-мерного пространства.

Для анализа расстояния между точками n-мерного пространства предлагается использовать визуальное отображение этих точек. В начале осуществляется проецирование исходного множества точек на одно из трехмерных пространств. При этом:

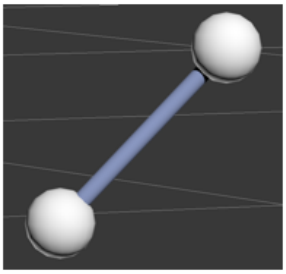
- Многомерная точка pi проецируется в сферу Si.

$$p_i = (p_i^1, p_i^2, \ldots, p_i^n) \xrightarrow[SphC]{p_i^{l_1}, p_i^{l_2}, p_i^{l_3}, SphR}$$



- Если расстояние между точками n-мерного пространства p1 и p2 меньше порогового расстояния d, задаваемого аналитиком в интерактивном режиме, то строится цилиндр, соединяющий сферы S1 и S2.
- Цвет цилиндра моделирует расстояние между точками p1 и p2 от красного (малое расстояния) до синего (большое расстояние).

$$p_{i_1}, p_{i_2}, d\colon \rho(p_{i_1}, p_{i_2}) \le d \xrightarrow[CylC]{CylR}$$



Затем выполняется графическое проецирование сфер и цилиндров на картинную плоскость с последующим их визуальным анализом. Результирующая совокупность сфер и цилиндров образует пространственную сцену с заданной геометрией и оптическими (цветовыми) характеристиками.

Таким образом, визуальный анализ пространственной сцены позволит судить о расстоянии между исходными многомерными точками. В процессе решения задачи анализа предлагается задание в начале исходного большого значения d, а затем проводить его уменьшение и выделять подмножества многомерных точек в зависимости от получаемого изображения на картинной плоскости. Следует отметить, что в рамках этого подхода аналитик в процессе анализа данных не пассивно созерцает пространственную сцену, а имеет возможность интерактивного взаимодействия с ней.

4. Области применения визуальной аналитики

4.1. Обнаружение связей (и влияния) между параметрами объектов данных

Интерпретация параметров объектов данных и метрик в N-мерном пространстве (размерность определяется числом исследуемых параметров) для оценки взаимного влияния и определения их корреляций.

4.2. Автоматическое выявление аномалий при выполнении задач обработки, анализа и моделирования

Интерпретация данных в N-мерном пространстве (размерность определяется числом исследуемых параметров) и проецирование в 3х-мерное пространство позволит производить кластеризацию объектов (по заданным параметрам и метрикам) с целью обнаружения объектов с нетипичным набором значений заданных параметров, т.н., аномальные объекты. В случае системы управления распределенными потоками заданий такими объектами данных выступают вычислительные задания (tasks) и задачи (jobs), описывающие процессы обработки данных в соответствующих специализированных системах (ProdSys2 и PanDA). Сбор необходимой информации о процессах обработки (параметры объектов и метрики) производится при формировании заданий и задач, и их последующего выполнения (возможен учет параметров и метрик, описывающих текущее состояние вычислительной среды и используемых вычислительных ресурсов).

4.3. Определение причин неэффективности процессов обработки данных

Определение последовательности этапов при которых произошел сбой, задержка или ошибка в процессе обработки данных, и выявление набора значений заданных параметров в начальном этапе, послуживших причиной сбоя (на основе методов, разработанных в п.4.1).

4.4. Определение популярности данных для динамического управления данными

Популярность данных определяется количеством обращений задач анализа к наборам данных и количеством запросов на дополнительное копирование данных. При увеличении количества обращений к данным требуется автоматически создавать дополнительные копии, географически “привязанные” к месту наибольшей востребованности.

Исследования, проведенные с помощью системы мониторирования, показали, что “популярность” данных резко уменьшается примерно через 45 дней, что позволяет принять решение об удалении копий невостребованных наборов данных с жестких дисков и их переноса для хранения на ленту.

Методы визуальной аналитики позволят кластеризовать данные по географическому месту хранения (ВЦ) и их востребованности в зависимости от времени. Это позволит оптимизировать требования о создании/удалении дополнительных/излишних копий данных.

5. Апробация предлагаемого подхода

5.1. Визуальная аналитика статистических данных о вычислительных задачах системы ProdSys2/PanDA

5.1.1. Постановка задачи

Разработка программных средств визуализации и методики их использования для кластерного анализа кортежей параметров вычислительных задач системы PanDA. Визуальный анализ позволит выявлять схожие (подобные) задачи, а также аномальные задачи, при этом определять за счет каких параметров эта аномальность имеет место.

Разработка программных средств визуализации и методики их использования для анализа длительностей вычислительных задач системы PanDA. Визуальный анализ позволит определить влияние определенного набора параметров (т.н., базовый набор параметров, который может быть расширен в дальнейшем) на продолжительность времени обработки вычислительных задач, и выявление набора/диапазона значений отдельных параметров, индицирующих увеличение времени выполнения вычислительных задач (т.н., отклонение от среднего времени, затрагивает задачи с временем выполнения превышающие 3σ, предполагается что это около 7.5% от числа всех задач).

Разработка программных средств визуализации и методики их использования для анализа популярности данных в зависимости от времени. Визуальный анализ позволит выявить увеличение/уменьшение количества обращений к данным в зависимости от времени, обеспечив, тем самым, удобный визуальный метод принятия решений о динамическом управлении распределения реплик данных.

5.1.2. Ожидаемый результат

Построение проекционных графических изображений N-мерных геометрических интерпретаций параметров задач, представление в виде 2D гистограмм количества задач (ось Y), сгруппированных по длительности/времени их выполнения (ось X), и выявление группы задач с “увеличенной” длительностью выполнения.

5.1.3. Инструментарий

Для решения поставленной исследовательской задачи определены ключевые параметры описания вычислительных задач:

- Время выполнения задачи (<duration> = endtime - starttime)
- Наименование группы задач (gShare)
- Центр запуска и обработки задачи (nucleus)
- Количество обрабатываемых событий (nEvents)
- Дополнительный набор параметров (расширение базового набора):
  - Название стадии анализа/обработки данных (processingType); объем входных данных для задачи (inputFileBytes); тип входных данных (inputFileType); объем выходных данных задачи (outputFileBytes); исходный приоритет задачи (assignedPriority); процессорное время обработки одного события (cpuTimePerEvent); аппаратная архитектура, на которой выполняются вычисления для данной задачи (cmtConfig); количество ядер (actualCoreCount); релиз программного обеспечения (atlasRelease); эффективность процессора на ядро (CPU eff per core); средний размер страниц памяти, выделенных процессу операционной системой и в

настоящее время находящихся в ОЗУ (avgRSS); средняя доля общей памяти, используемой процессором (avgPSS); средний размер выделенной виртуальной памяти (avgVMEM); максимальный размер страниц памяти, выделенных процессору операционной системой (maxRSS); максимальная доля общей памяти, используемой процессором (maxPSS); максимальный размер выделенной виртуальной памяти (maxVMEM)

- Индицирующие параметры:
  - Шаг перезапуска задачи (attemptNr); коды ошибок (brokerageErrorCode, ddmErrorCode, exeErrorCode, jobDispatcherErrorCode, pilotErrorCode, supErrorCode, taskBufferErrorCode)

Формат представления исходных данных:

- Объединение параметров описания задач в группы
- Представление входных данных в виде матриц, соответствующих группе параметров, где строки (rows) соответствует записям о задачах, а столбцы (columns) соответствуют параметрам определенной группы:
  - $D_{n\times1}$ - матрица с длительностями выполнения задач, где n - количество задач;
  - $P_{n\times3}$ - матрица с базовым набором параметров задач (gShare, nucleus, nEvents);
  - $E_{n\times16}$ - матрица с дополнительным набором параметров;
  - $I_{n\times8}$ - матрица с индицирующими параметрами задач (attemptNr, errorCodes).

Источник данных:

- Инфраструктура ElasticSearch в Университете Чикаго [18]
- Индексы “jobs_archive_*”
  - Условия поиска и выгрузки данных
    - Допустимые статусы задач: jobStatus IN (“finished”, “failed”);
    - Источник задач: prodSourceLabel = ‘managed’
    - Тип обрабатываемых данных и этап обработки: REGEXP_LIKE (jobName, “^mc(.*\.){3}simul\..*”)

5.2. Визуальная аналитика статистических данных о выполнении заданий ProdSys2/PanDA

Данная исследовательская задача подразумевает расширение поставленной задачи п.5.1 по отношению к вычислительным заданиям ProdSys2 и решение исходит из полученных результатов исследовательской задачи п.5.1, т.к. подразумевает аналогичный подход, но с учетом специфики рассматриваемых объектов данных - вычислительные задания, на основе которых формируются наборы вычислительных задач.

6. Основные этапы выполнения проекта

1. Апробация подхода по применению визуальной аналитики на примере использования кластерного анализа кортежей параметров вычислительных задач PanDA и оценки распределения длительности выполнения вычислительных задач.
2. Расширение разработанного подхода применительно к вычислительным заданиям ProdSys2.
3. Интеграция разработанных прототипов средств визуализации и

аналитики в инфраструктуру контроля и мониторинга системы ProdSys2/PanDA.

4. Оценка модификации существующего процесса контроля и мониторинга ProdSys2/PanDA при применении подхода визуальной аналитики.

7. Ожидаемые результаты проекта

В результате выполнения проекта будет разработана визуально-аналитическая система для мониторинга распределенных систем обработки данных. Разработанная система будет представлять собой расширенную аналитическую службу существующей системы мониторинга и контроля эксперимента ATLAS. Посредством разработанной системы функциональные возможности мониторинга будут существенно расширены, позволяя моделировать, прогнозировать дальнейший ход выполнения эксперимента. Визуальная аналитика ляжет в основу системы поддержки принятия решений, и стратегического планирования.

Сотрудничество с экспериментом ATLAS на LHC, наличие доступа к данным эксперимента и демонстрация созданного решения и прототипа на реально работающей системе обработки данных, обеспечит уникальный испытательный полигон для отработки технологий аналитических исследований и применения методов визуальной аналитики и позволит данному проекту быть в ряду важнейших мировых разработок в данной области.

Результаты проекта будут востребованы при создании ПО коллайдера NICA (ОИЯИ, Дубна), для этапа высокой светимости LHC (HL-LHC), а также для визуализации научной информации на таких мегаустановках как XFEL и FAIR.

8. Основные участники пилотного проекта

В пилотном проекте принимают участие международная коллаборация ATLAS, НИЯУ МИФИ и Российские научные центры.

Список университетов и научных центров:

- Национальный исследовательский ядерный университет “МИФИ”
  - Лаборатория научной визуализации
  - Кафедра анализа конкурентных систем
  - Группа МИФИ в эксперименте АТЛАС
- Национальный исследовательский центр “Курчатовский институт”
  - Лаборатория Технологий Больших Данных
- Национальный исследовательский Томский политехнический университет
- Объединенный Институт Ядерных Исследований
  - Лаборатория Информационных Технологий
- Брукхейвенская Национальная Лаборатория
- Университет Айовы
- Университет Чикаго
- Университет Техаса в Арлингтоне
- Европейский Центр Ядерных Исследований (ЦЕРН)

9. Работа со студентами и аспирантами и преподавательская деятельность

Одной из задач проекта является работа со студентами и аспирантами, в том числе подготовка бакалавров, магистров и аспирантов, владеющих современным инструментарием научной визуализации и работы с данными физического эксперимента. В НИЯУ МИФИ преподаются курсы “Визуальная аналитика” и “Научная визуализация”, на основе результатов проекта специальные курсы

будут созданы для магистрантов по специальности физика частиц и ядерная физика, и для магистрантов по специальности системотехника. Кроме того, Университет “Дубна” и Институт Кибернетики ТПУ выразили заинтересованность в создании совместных курсов по тематике проекта (в Университете “Дубна” создан учебный курс для подготовки специалистов для работы на коллайдере NICA. ТПУ активно участвует в научной программе в области физики частиц: эксперимент COMPASS на суперпротонном синхротроне (SPS, ЦЕРН) и эксперименты ATLAS и CMS на LHC).

10. Информационная поддержка

Информационная поддержка проекта осуществляется журналом "Научная визуализация" [19], а также порталами эксперимента ATLAS в ЦЕРН [20], Лаборатории Больших Данных НИЦ КИ [21] и ЛИТ ОИЯИ [22].

Список литературы

1. The ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider”, Journal of Instrumentation, vol. 3, S08003, 2008.
2. LHC - Large Hadron Collider, http://lhc.web.cern.ch/lhc
3. 26th International Symposium on Nuclear Electronics & Computing - NEC’2017, http://indico.jinr.ru/conferenceDisplay.py?confId=151
4. S.Padolski, T.Korchuganova, T.Wenaus, M.Grigorieva, A.Alexeev, M.Titov, A.Klimentov, "Data visualization and representation in ATLAS BigPanDA monitoring", Scientific Visualization, 2018.
5. J.Thomas, K.Cook, "Illuminating the Path: The Research and Development Agenda for Visual Analytics", IEEE Computer Society, 2005.
6. D.Popov, I.Milman, V.Pilyugin, A.Pasko, "Visual Analytics of Multidimensional Dynamic Data with a Financial Case Study", Data Analytics and Management in Data Intensive Domains, Springer International Publishing, pp. 237--247, 2017.
7. M.Borodin, K.De, J.Garcia Navarro, D.Golubkov, A.Klimentov, T.Maeno, A.Vaniachine, “Scaling up ATLAS production system for the LHC Run 2 and beyond : project ProdSys2”, Journal of Physics: Conference Series, vol. 664, no. 6, 2015.
8. A.Klimentov et al., "Migration of ATLAS PanDA to CERN", Journal of Physics: Conference Series, vol. 219, no. 6, 2010.
9. WLCG - Worldwide LHC Computing Grid, http://wlcg.web.cern.ch
10. M.Titov, G.Zaruba, A.Klimentov, and K.De, “A probabilistic analysis of data popularity in ATLAS data caching”, Journal of Physics: Conference Series, vol. 396, no. 3, 2012.
11. M.Titov, M.Gubin, A.Klimentov, F.Barreiro, M.Borodin, D.Golubkov, "Predictive analytics as an essential mechanism for situational awareness at the ATLAS Production System", The 26th International Symposium on Nuclear Electronics and Computing (NEC), CEUR Workshop Proceedings, vol. 2023, pp. 61--67, 2017.
12. Y.Livnat, J.Agutter, S.Moon, S.Foresti, "Visual correlation for situational awareness", IEEE Symposium on Information Visualization (INFOVIS), pp. 95--102, 2005.
13. D.Mavris, O.Pinon, D.Fullmer, "Systems design and modeling: A visual analytics approach", Proceedings of the 27th International Congress of the Aeronautical Sciences (ICAS), 2010.

14. SAS the power to know, [Online]. Available: http://www.sas.com/en_us/home.html [accessed on 15.03.2018].
15. Шаропин К.А. [и др.] Визуализация медицинских данных на базе пакета NovoSpark [Журнал] // Известия Южного федерального университета. Технические науки. – 2010. – том 109. – стр. 242-249.
16. J.Woollen, "A Visual Approach to Improving the Experience of Health Information for Vulnerable Individuals", PhD Thesis, Columbia University Academic Commons, 2018.
17. Пилюгин В.В. Научная визуализация как метод анализа научных данных / В.В. Пилюгин [и др.] // Научная визуализация. – 2012. – том 4. – стр. 56-70.
18. http://atlas-kibana.mwt2.org:5601/app/kibana
19. http://sv-journal.org
20. http://atlas.cern
21. http://bigdatalab.nrcki.ru
22. http://lit.jinr.ru