Дайджест литературы и статей

Тема: *«Исследование методов и подходов к построению распределённых систем для обработки больших объёмов данных и разработка архитектуры распределённой аналитической платформы для проекта sdLitica»*

---

[1] Designing Distributed Systems: Patterns and Paradigms for Scalable, Reliable Services

Книга посвящена формулированию паттернов проектирования распределённых систем. Рассматривается эволюция подходов к разработке и принятые практики в современном мире. Аналогично формализации алгоритмического программирования, а затем появлению паттернов в объектно-ориентированном программировании, автор систематизирует популярные и необходимые паттерны, практики и обобщённые переиспользуемые компоненты в сфере сложных распределённых вычислительных систем. В трёх частях последовательно излагаются одноузловые паттерны проектирования, паттерны проектирования обслуживающих систем и паттерны проектирования систем пакетных вычислений. В результате создаётся необходимый минимум компетенций для понимания устройства большинства современных крупных систем и формируется багаж общепринятых лучших практик и подходов к разработке сложных составных продуктов.

> Burns, Brendan. Designing Distributed Systems: Patterns and Paradigms for Scalable, Reliable Services. "O'Reilly Media, Inc.", 2018. URL: https://books.google.ru/books?vid=ISBN9781491983645

---

[2] Distributed systems: principles and paradigms

Несмотря на то, что первое издание этой книги датируется 2003 годом, эта работа является фундаментальным курсом по распределенным системам. Книгу логически образуют две части — принципы и парадигмы. В первой главе дается введение в тему. Затем следует семь глав, описывающих отдельные принципы, которые авторы сочли особенно важными: связь, процессы, именование, синхронизация, непротиворечивость и репликация, защита от сбоев и безопасность. Особое внимание в книге уделено World Wide Web, развитие которой и послужило толчком к резкому повышению интереса к распределенным системам. Авторы приводят последовательное и детальное изложение теории, которая в свою очередь сопровождается примерами реально действующих систем.

> Tanenbaum, Andrew S., and Maarten Van Steen. Distributed systems: principles and paradigms. Prentice-Hall, 2007. URL: https://books.google.ru/books?vid=ISBN0132392275

---

[3] Выбор технологических решений для разработки программного обеспечения распределенных информационных систем

В статье рассматривается выбор технологических решений на примере разрабатываемой цифровой вычислительной веб-платформы Российской академии образования для обеспечения информационной поддержки деятельности психологов. Для разрабатываемой системы были введены критерии оценки программных технологий, которые учитывают особенности функционирования и жизненного цикла продукта, на конкретном примере показан выбор соответствующих технологических решений. Приведена система, реализующая программу обучения выбранных технологий.
Однако перечень рассматриваемых авторами технологий для backend разработки не является исчерпывающим хотя бы в силу того, что в сравнение не вошли один из самых популярных фреймворков для языков семейства JVM — Spring и быстро набирающий популярность в области разработки высоконагруженных распределённых систем язык Go.

> Ильин Дмитрий Юрьевич, Никульчев Евгений Витальевич, Колясников Павел Владимирович. Выбор технологических решений для разработки программного обеспечения распределенных информационных систем // Современные информационные технологии и ИТ-образование. 2018. №2. doi: https://doi.org/10.25559/SITITO.14.201802.344-354

---

[4] The evolution of distributed systems towards microservices architecture

Статья рассматривает, как распределенные системы эволюционировали от традиционной монолитной модели клиент-сервер к сервис-ориентированной, а после к недавно предложенной микросервисной архитектуре. Производится обзор всех архитектур, приводятся соответсвующие известные работы и обоснования того, почему они должны были появиться и развиваться на конкретном этапе. Также предоставляется сравнительный анализ всех архитектур и делаются выводы о соотвествии их возможностей с требованиями реалий.

> Salah, Tasneem, M. Jamal Zemerly, Chan Yeob Yeun, Mahmoud Al-Qutayri, and Yousof Al-Hammadi. "The evolution of distributed systems towards microservices architecture." In 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 318-325. IEEE, 2016. doi: https://doi.org/10.1109/icitst.2016.7856721

---

[5] Differences between model-driven development of service-oriented and microservice architecture

В этой статье исследуется возможность применения годами накопленного опыта в сфере Model-Driven Development (MDD) для сервис-ориентированной архитектуры (SOA) к новому популярному подходу микросервисной архитектуры (MSA). Таким образом, авторы определяют

концептуальные и практические различия между SOA и MSA и классифицируют их на основе иерархической схемы. Исходя из выявленных различий, мы делаем вывод о влиянии MSA на MDD и обсуждаем их с учетом существующей совокупности знаний о MDD SOA. Таким образом, мы даем первоначальный обзор различий между SOA и MSA, а также последствия в определенных областях MDD, которые следует учитывать при адаптации сервис-ориентированного MDD к MSA

> Rademacher, Florian, Sabine Sachweh, and Albert Zündorf. "Differences between model-driven development of service-oriented and microservice architecture." In 2017 IEEE International Conference on Software Architecture Workshops (ICSAW), pp. 38-45. IEEE, 2017. doi: https://doi.org/10.1109/ICSAW.2017.32

---

[6] Towards Integrating Microservices with Adaptable Enterprise Architecture

Авторы статьи расширяют методологии и модели Enterprise Architecture (EA), которые охватывают высокую степень неоднородности и распределения в современных программных системах, чтобы поддерживать цифровую трансформацию и связанные информационные системы с микрогранулярными архитектурами. Основная цель — обеспечить гибкость трансформации как ИТ, так и бизнес-возможностей в рамках адаптируемых цифровых корпоративных архитектур. В данной исследовательской работе исследуются механизмы интеграции микросервисных архитектур (MSA) путем расширения исходных эталонных моделей архитектуры предприятия элементами для более гибких архитектурных мета-моделей и мини-описаний EA.

> Bogner, Justus, and Alfred Zimmermann. "Towards integrating microservices with adaptable enterprise architecture." In 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW), pp. 1-6. IEEE, 2016. doi: https://doi.org/10.1109/EDOCW.2016.7584392

[7] Практика создания распределённых систем

В работе рассматриваются отдельные аспекты проблем построения распределенных систем. На основе системного подхода к исследованию проанализированы виды информации, значимые в распределенных системах, а также способы взаимодействия элементов распределенной системы, представление и обработка данных в процессе взаимодействия, сбои и безопасность как состояние распределенной системы. Сделан вывод о необходимости учета ряда факторов, влияющих на распределенные системы на этапах анализа и проектирования.

> Курбацкий, А.Н. Практика создания распределенных систем / А.Н. Курбацкий, С.Е. Довнар // Актуальные проблемы науки XXI века. 2017. URL : http://elibrary.miu.by/journals!/item.science-xxi/issue.6/article.1.html

[8] The System Design Primer

Созданный, поддерживаемый и развиваемый сообществом сборник информации и различных ресурсов об архитектуре масштабируемых систем. Репозиторий освещает такие темы, как производительность, доступность и надёжность систем, а также рассматриваются отдельные составные части систем — клиенты и сервера, реляционные и NoSQL базы данных, балансировщики нагрузки, кеши, API и коммуникационные протоколы. Также приводятся примеры проектирования архитектур популярных компаний, таких как Twitter или Uber.

> The System Design Primer // GitHub repository // Access: 28.10.2020. URL: https://github.com/donnemartin/system-design-primer

[9] Time Series Similarity Search for Streaming Data in Distributed Systems

В этой статье авторы проводят практическое исследование и демонстрацию поиска схожести временных рядов в современных платформах распределенной обработки данных для потоковых данных. После тщательного изучения литературы авторы реализуют гибкое приложение для поиска схожести в Apache Flink, которое включает в себя наиболее часто используемые измерения расстояний: евклидово расстояние и динамическую трансформацию временной шкалы. Для эффективного и точного поиска схожести оцениваются методы нормализации и сокращения, разработанные для обработки на одном компьютере, и демонстрируется возможность адаптации и использовании для этих распределенных платформ. Окончательная реализация способна отслеживать множество временных рядов в реальном времени и параллельно. Кроме того, демонстрируется, что количество требуемых параметров может быть уменьшено и оптимально получено из свойств данных. Проводятся замеры производительности на данных электрокардиограммы в кластере с несколькими узлами. В результате достигается среднее время отклика менее 0,1 мс для окон данных по 2 с, что позволяет быстро реагировать на схожие последовательности.

> Ziehn, Ariane, Marcela Charfuelan, Holmer Hemsen, and Volker Markl. "Time Series Similarity Search for Streaming Data in Distributed Systems." In EDBT/ICDT Workshops. 2019. URL: http://www.redaktion.tu-berlin.de/fileadmin/fg131/dima-feed/TimeSeriesSimilaritySearch_Darli-2019_crv.pdf

---

[10] Big data time series forecasting based on nearest neighbours distributed computing with Spark

В данной работе представлен новый подход к прогнозированию больших данных, основанный на алгоритме k-взвешенных ближайших соседей. Такой алгоритм был разработан для распределенных вычислений в рамках Apache Spark. Объясняется каждый этап алгоритма, а также способы получения необходимых оптимальных значений входных параметров. Для тестирования разработанного алгоритма использовался временной ряд больших данных о потреблении энергии в Испании.

Оценка точности прогноза показала замечательные результаты. Кроме того, обсуждается оптимальная конфигурация кластера Spark. Наконец, был проведен анализ масштабируемости алгоритма, в результате которого был сделан вывод, что предложенный алгоритм хорошо подходит для сферы больших данных.

> Talavera-Llames, Ricardo & Pérez-Chacón, Rubén & Troncoso, Alicia & Martínez-Álvarez, Francisco. Big data time series forecasting based on nearest neighbours distributed computing with Spark. Knowledge-Based Systems. 2018. doi: https://doi.org/10.1016/j.knosys.2018.07.026

---

[11] Distributed Time Series Data Management and Analysis

В NTT Network Innovation Laboratories изучаются сложные схемы управления сетью, основанные на анализе сетевых данных. В этой статье авторы сосредотачиваются на схеме распределенного анализа данных для больших объемов сетевых данных. С помощью этой схемы можно уменьшить объем сетевых данных, передаваемых между центрами обработки данных для анализа, что также снизит затраты.

> Kimihiro Mizutani, Takeru Inoue, Toru Mano, Hisashi Nagata, and Osamu Akashi. Distributed Time Series Data Management and Analysis. NTT Technical Review, Feature Articles: New Generation Network Platform and Attractive Network Services. 14-3. 2016. URL: https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201603fa5.html

---

[12] A Periodicity-based Parallel Time Series Prediction Algorithm in Cloud Computing Environments

В этой статье предлагается алгоритм параллельного прогнозирования временных рядов на основе периодичности (PPTSP) для

крупномасштабных данных временных рядов, который реализуется в среде облачных вычислений Apache Spark. Для эффективной обработки массивных наборов исторических данных представлен алгоритм сжатия и абстракции данных временных рядов (TSDCA), который может уменьшить масштаб данных, а также с высокой точностью извлекать характеристики данных. Основываясь на этом, авторы предлагают алгоритм распознавания периодических образов многослойных временных рядов (MTSPPR) с использованием метода анализа спектра Фурье (FSA). Кроме того, предлагается алгоритм прогнозирования временных рядов на основе периодичности (PTSP).

> Chen, Jianguo, Kenli Li, Huigui Rong, Kashif Bilal, Keqin Li, and S. Yu Philip. "A periodicity-based parallel time series prediction algorithm in cloud computing environments." Information Sciences 496 (2019): 506-537. doi: https://doi.org/10.1016/j.ins.2018.06.045

---

[13] Distributed Time Series Analytics

В этой работе рассматриваются важные вопросы, относящиеся к методам масштабируемой и распределенной аналитики для больших данных временных рядов. Конкретно, этот тезис сосредоточен вокруг следующих трех тем: управление и построение запросов к model-view временным рядам, вычисление корреляций в потоковых временных рядах, обучение на зашумленных и нестационарных данных.

> Guo, Tian. Distributed Time Series Analytics. No. THESIS. EPFL, 2017. doi: https://doi.org/10.5075/epfl-thesis-7395

[14] Distributed Algorithms to Find Similar TimeSeries

Авторы статьи используют финансовые и сейсмические данные, чтобы показать, как два современных алгоритма создают индексы и отвечают на запросы схожести временных рядов с помощью Spark. Посетители демо-версии смогут выбрать временной ряд запроса, увидеть, как каждый алгоритм аппроксимирует ближайших соседей, и сравнить время в параллельной среде.

> Oleksandra Levchenko, Boyan Kolev, Djamel-Edine Yagoubi, Dennis Shasha, Themis Palpanas, et al.. Distributed Algorithms to Find Similar Time Series. ECML-PKDD 2019 - European Conference on Machine Learning and Knowledge Discovery in Databases, Sep 2019, Wurtzbourg, Germany. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265726

---

[15] A Comparison of Distributed Stream Processing Systems for Time Series Analysis

Учитывая огромное количество систем обработки данных, доступных сегодня, в этой статье авторы стремятся идентифицировать, выбрать и оценить системы, наиболее подходящие для использования при проведении анализа временных рядов. Опубликованные исследования производительности используются для сравнения нескольких систем с открытым исходным кодом, а две системы дополнительно выбираются для качественного сравнения и оценки в отношении определённой задачи анализа временных рядов. В качестве тестового сценария реализован дискретный фильтр Калмана для прогнозирования цены закрытия на данных фондового рынка в режиме реального времени. Рассматривается охват базовой функциональности, а расширенная функциональность оценивается с использованием нескольких критериев качественного сравнения.

> Gehring, Melissa, Marcela Charfuelan, and Volker Markl. "A comparison of distributed stream processing systems for time series analysis." BTW 2019–Workshopband (2019). doi: https://doi.org/10.18420/btw2019-ws-21

---

[16] Massively Distributed Time Series Indexing and Querying

Авторы предлагают решение для параллельного индексирования, которое изящно масштабируется для набора из миллиардов временных рядов, и стратегию параллельной обработки запросов, которая при наличии набора запросов эффективно использует построенный индекс.

> Yagoubi, Djamel-Edine, Reza Akbarinia, Florent Masseglia, and Themis Palpanas. "Massively distributed time series indexing and querying." IEEE Transactions on Knowledge and Data Engineering 32, no. 1 (2018): 108-120. doi: https://doi.org/10.1109/TKDE.2018.2880215

[17] Fast Distributed Correlation Discovery Over StreamingTime-Series Data

Ключевой проблемой при обнаружении корреляций временных рядов в реальном времени является то, что количество пар, которые должны быть проанализированы, растет квадратично по отношению к количеству временных рядов, что приводит к квадратичному увеличению как затрат на вычисления, так и затрат на связь между узлами кластера в распределенной системе. Чтобы решить эту проблему, авторы предлагают структуру под названием AEGIS. AEGIS использует хорошо зарекомендовавшие себя статистические свойства, чтобы резко сократить количество пар временных рядов, которые необходимо оценивать для обнаружения интересных корреляций. Приведённые в статье экспериментальные оценки реальных и синтетических наборов данных устанавливают эффективность AEGIS по сравнению с исходными аналогами.

> Tian Guo, Saket Sathe, and Karl Aberer. 2015. Fast Distributed Correlation Discovery Over Streaming Time-Series Data. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15). Association for Computing Machinery, New York, NY, USA, 1161–1170. doi: https://doi.org/10.1145/2806416.2806440

[18] Distributed and Scalable Platform for Collaborative Analysis of Massive Time Series Data Sets

В статье предлагается распределенная, масштабируемая, безопасная и высокопроизводительная архитектура, которая позволяет группе исследователей работать над общей базой знаний, развернутой в сети, и

аннотировать шаблоны, предотвращая при этом потерю данных из-за пересекающегося редактирования или несанкционированных изменений.

> Duarte, Ed & Gomes, Diogo & Aguiar, Rui. (2019). Distributed and Scalable Platform for Collaborative Analysis of Massive Time Series Data Sets. 41-52. doi: https://doi.org/10.5220/0007834700410052

[19] Distributed ARIMA Models for Ultra-long Time Series

В этой статье разрабатывается новая структура распределенного прогнозирования для решения проблем, связанных с прогнозированием сверхдлинных временных рядов, с использованием стандартной инфраструктуры MapReduce. Предлагаемый подход облегчает распределенное прогнозирование временных рядов путем комбинирования локальных оценок моделей ARIMA (AutoRegressive Integrated Moving Average), полученных от рабочих узлов, и минимизации глобальной функции потерь. Таким образом, вместо нереалистичного предположения, что процесс генерации данных (DGP) большого временного ряда остается неизменным, авторы делаем предположения только на DGP подсерии, охватывающей более короткие периоды времени. Также исследуется производительность предложенных распределенных моделей ARIMA на наборе данных о спросе на электроэнергию. По сравнению с моделями ARIMA предлагаемый подход приводит к значительному повышению точности прогнозов и вычислительной эффективности как точечных прогнозов, так и интервалов прогнозирования, особенно для более длинных горизонтов прогнозов.

> Wang, Xiaoqian, Yanfei Kang, Rob J. Hyndman, and Feng Li. "Distributed ARIMA Models for Ultra-long Time Series." arXiv preprint arXiv:2007.09577 (2020). URL: https://arxiv.org/abs/2007.09577