УДК 004.021:004.75

ИНТЕГРАЦИЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В АВТОМОБИЛЬНОЙ ПРОМЫШЛЕННОСТИ

С.А. Мигалевич
Магистр технических наук, начальник центра информатизации и инновационных разработок
migalevich@bsuir.by

А.Н. Марков
Магистр технических наук, заместитель начальника центра информатизации и инновационных разработок
a.n.markov@bsuir.by

Д.Г. Ершов
Студент Белорусского государственного университета информатики и радиоэлектроники
d.ershov@bsuir.by

С.А. Мигалевич
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: вибродиагностика, разработка метода вейвлет–анализа изделий машиностроения.

А.Н. Марков
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: вычислительные системы, облачные вычисления (CLOUD COMPUTING), распределенные вычислительные системы, балансировка нагрузки вычислительных систем (load balancing).

Д.Г. Ершов
Студент Белорусского государственного университета информатики и радиоэлектроники факультета компьютерного проектирования.

Аннотация. По мере того как технология Интернета вещей (*IoV*) становится все более важным трендом для будущего транспорта, разработка крупных систем *IoV* становится критической задачей, направленной на обработку больших данных, загружаемых автопарками, и предоставление сервисов на основе данных. Данные *IoV*, особенно статусы транспортных средств с высокой частотой (например, местоположение, параметры двигателя), характеризуются большим объемом с низкой плотностью значений и низким качеством данных. Такие характеристики создают проблемы для разработки приложений в реальном времени на основе таких данных. В этой статье мы рассматриваем проблемы проектирования масштабной системы *IoV*, описывая *CarStream*, промышленную систему обработки больших данных для услуг аренды автомобилей с водителем. Подключенная к более чем 30 000 автомобилям, *CarStream* собирает и обрабатывает несколько типов данных о вождении, включая статус автомобиля, активность водителя и информацию о поездках пассажиров. На основе собранных данных предоставляются несколько сервисов. *CarStream* была задействована и поддерживается в промышленном использовании в течение трех лет, собрав более 40 терабайт данных о вождении. В данной статье рассмотрен опыт проектирования *CarStream* на основе потоков данных о вождении крупномасштабного характера, а также уроками, извлеченными из процесса решения проблем при проектировании и поддержке *CarStream*.

Ключевые слова: большие данные, автомобильная промышленность, отслеживание рисков, управление знаниями, автоматизация эксплуатации.

Введение. В последние годы технология Интернета вещей (*IoT*) стала важной областью исследований и применения. В качестве одного из основных направлений *IoT* Интернет вещей для транспорта (*IoV*) привлек большое внимание исследователей и промышленности. Недавно облачные технологии *IoV* получили выгоду от быстрого развития мобильных сетей и технологий обработки больших данных. В отличие от традиционных технологий сетей транспортных средств, которые фокусируются на взаимодействии транспортных средств (*V2V*) и сетях транспортных средств, в типичном облачном сценарии *IoV* транспортные средства подключены к облачному центру данных и передают статусы транспортных средств в центр через беспроводные коммуникации. Облако собирает и анализирует переданные данные и отправляет обратно на транспортные средства дополнительную информацию со значением. Аналогично другим приложениям *IoT*, данные о транспортных средствах обычно организованы в виде потока. Хотя каждое транспортное средство загружает небольшой поток данных, в облаке объединяется большой поток из–за как высокой частоты, так и большого масштаба флота. Поэтому основным требованием в этом облачном сценарии *IoV* является обработка потока больших данных о транспортных средствах в своевременном режиме (рисунок 1).

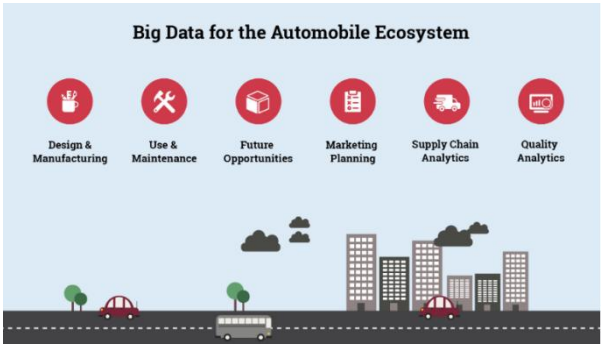

Рисунок 1. Большие данные в автомобильной промышленности

Существует несколько решений для проектирования системы *IoV*. Например, традиционная архитектура *OLTP* (*Online Transaction Processing*) использует зрелую и стабильную базу данных в качестве центра для развертывания сервисов. В такой архитектуре система собирает данные, загруженные транспортными средствами, и сохраняет их в базе данных, предоставляя дополнительные сервисы на основе аналитической способности подсистемы базы данных. Это решение просто, зрело и может быть очень надежным. Однако оно не масштабируется хорошо с ростом размера флота, потому что база данных легко может стать узким местом всей системы. Архитектура *OLAP* (*Online Analytical Processing*) часто используется для работы с крупномасштабными данными, в которой подсистема обработки данных, а не подсистема хранения данных (например, подсистема базы данных), предоставляет основную аналитическую способность. В такой системе нагрузка на вычисления в основном ложится на подсистему обработки данных, которая часто выполняет сложные запросы. Проектирование подсистемы обработки данных в стиле *OLAP* для сценариев *IoV* часто требует интеграции нескольких платформ, таких как *Storm* и *Spark Streaming* [1].

В данной статье рассмотрены задачи проектирования масштабной и высокопроизводительной системы аналитики больших данных для *IoV*, описывая *CarStream* – промышленную систему обработки данных для услуг аренды автомобилей с водителем. *CarStream* соединил более 30 000 автомобилей в более чем 60 городах и имеет несколько источников данных, включая данные о состоянии транспортного средства (такие как скорость, траектории, обороты двигателя в минуту (об/мин), оставшееся

количество бензина), активности водителя (например, начало обслуживания, прием пассажира) и заказы пассажиров. *CarStream* предоставляет несколько услуг в реальном времени на основе этих данных.

В частности, решается проблема масштабируемости, оснащая *CarStream* способностью к распределенной обработке и хранению данных. Затем используется подсистема обработки потоков для предварительной обработки данных на лету, чтобы решить проблемы низкого качества данных и низкой плотности значения. Данное решение достигает более высокой производительности за счет большего объема хранения. Также ускоряется работа *CarStream* дальше, интегрируя подсистему кэширования в памяти. Наконец, разрабатывается трехуровневую систему мониторинга для *CarStream* для обеспечения высокой надежности и управляемости. Система мониторинга обеспечивает всестороннее представление о том, как работает вся система, включая кластерный уровень, уровень вычислительной платформы и уровень приложения. Эта подсистема помогает в реальном времени определять состояние здоровья *CarStream* и также предоставляет ценную информацию о системе разработчикам, непрерывно улучшающим надежность и производительность системы. *CarStream* был развернут и поддерживается на протяжении трех лет в промышленном использовании, в процессе которого продолжается совершенствование и обновление системы, а также развиваются новые приложения.

Сбор больших данных. При обсуждении сбора данных для данного анализа вглубь больших данных вовлекаются различные методы сбора информации. Например, в автомобильной промышленности используются разнообразные инструменты, такие как GPS, датчики и камеры, установленные в транспортных средствах. Все эти компоненты генерируют структурированные, неструктурированные и полуструктурированные данные (всего 5–10 процентов всех данных). Для обработки и анализа данных, полученных из этих источников. Полученные данные обрабатываются и объединяются для предоставления различных услуг, таких как прогнозирование движения автомобилей для компаний, предоставляющих телематические услуги, страховых компаний и агентств по прокату автомобилей. Эта информация используется для понимания спроса и предложения на продукты и услуги компаний, что позволяет им предоставлять клиентам более индивидуализированный и персонализированный опыт. Использование таких данных создает новую экосистему, улучшающую взаимодействие с пользователями.

Нахождение универсального инструмента, подходящего для всех сценариев обработки данных, представляет собой значительную сложность. Многие предприятия постоянно стремятся создать эффективное хранилище и обработчик больших данных, который также был бы удобен для разработчиков. Однако одним из ключевых подходов, обеспечивающих более управляемую обработку, является разделение данных на более мелкие фрагменты и их параллельная обработка. Применение метода "разделяй и властвуй", широко используемого в информатике, эффективно и для работы с большими данными.

Предыстория *CarStream*. Концепция *IoV* (Интернет вещей в автомобиле) вытекает из *IoT* (Интернет вещей). В традиционном понимании, *IoV* описывает сеть транспортных средств, соединенных через *RFID* (радиочастотную идентификацию). Технология *V2V* (*Vehicle to Vehicle*) дополнительно обменивается данными о трафике и состоянии автомобилей, обеспечивая решения на местном уровне для безопасности и эффективности транспортировки.

С развитием мобильного интернета, традиционное *IoV* перешло к развертыванию сети широкой зоны, объединяющей внутренние и межавтомобильные сети. Датчики автомобилей подключены к электронному блоку управления через *CAN-BUS*, а облачная сеть автомобилей строится с подключением к серверам через мобильную сеть [2].

Система *IoV* позволяет собирать, анализировать и хранить данные автомобилей на серверах, которые также отправляют обратно автомобилям полезную информацию.

IoV стал одним из самых популярных приложений в *IoT*, однако крупные развертывания *IoV* редки из-за высоких затрат на обслуживание и развертывание. Приватные автомобильные службы, такие как *Uber* и *Lyft*, успешные примеры таких систем. В этой статье обсуждаются проблемы проектирования облачной системы *IoV* под названием *CarStream*, которая основана на китайской компании *UCAR*. Компания предоставляет аналогичные услуги, но с флотом, принадлежащим самой компании, и водителями, являющимися ее сотрудниками. Для сбора данных используется *OBD* (*Onboard Diagnostic*), который передает информацию на сервер через беспроводной модуль (рисунок 2).

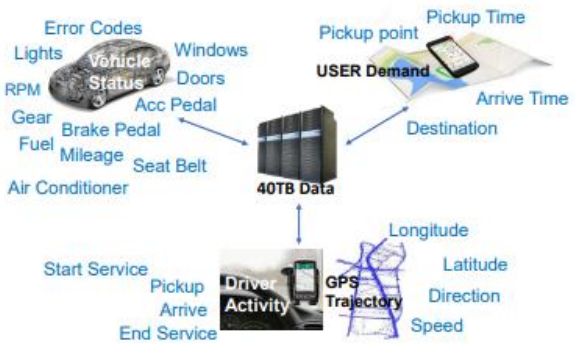

Рисунок 2. Устройство системы *CarStream*

CarStream обрабатывает данные о транспортных средствах и бизнесе в качестве сервиса на стороне сервера. На данный момент большая часть парка, более 30 000 транспортных средств, подключена к *CarStream*. Эти транспортные средства распределены по 60 различным городам Китая. Каждое транспортное средство загружает пакет данных на сервер каждые 1–3 секунды во время движения. Таким образом, серверная система должна обрабатывать почти 1 миллиард экземпляров данных в день. Объем парка расширился с 1500 до 30 000 за несколько месяцев; это расширение бизнеса создает запрос на горизонтальное масштабирование базовой платформы. В течение 3 лет *CarStream* собрало и обработало 40 ТБ данных о транспортных средствах.

Подсистема управления потоковыми данными. *CarStream* должен управлять терабайтами данных о поездках и предоставлять сервисы на основе анализа данных. Поток формируется большим количеством экземпляров данных, каждый из которых занимает сотни байт. *CarStream* ежедневно собирает почти миллиард экземпляров данных. Плотность полезной информации в таких данных очень низкая, так как из–за высокой частоты выборки можно извлечь лишь немного полезной информации. Кроме того, ценность данных быстро падает с течением времени, потому что большинство приложений, особенно те, что требуют временной чувствительности, в *IoV* более заинтересованы в новых данных.

Решение проблемы хранения данных в *IoV* с двух точек зрения. Первый аспект – это разделение данных. Выделяется приложения с высокими требованиями к реальному времени, чтобы далее выделить соответствующие «горячие» данные (например, текущий статус транспортного средства) отделяются «горячие» данные от «холодных» и помещаются в различные хранилища с разной производительностью. Второй аспект – это предварительная обработка. Сканирование больших наборов данных для базы данных затруднительно, но добавление данных (запись) может быть быстрым. Поэтому извлекается небольшой набор данных с более высокой плотностью ценности с помощью

потоковой обработки. Затем мы отделяем эту часть данных от сырых данных. Таким образом, подсистема хранения может обеспечить высокую пропускную способность для записи данных и высокую производительность для запросов, основанных на обработанных данных. Кроме того, мы помещаем некоторые «горячие» данные в кэш в памяти для обеспечения доступа к данным в реальном времени [3].

Для создания подсистемы управления данными мы использовали три различных платформы хранения в *CarStream*, включая кэширование в памяти, реляционные базы данных и хранилища больших данных *NoSQL*. Существует множество вариантов для каждого типа хранилища. Многие *NoSQL* базы данных, такие как *Cassandra*, *HBase* и *MongoDB*, могут быть использованы. На практике мы используем *HBase* в качестве архивного хранилища для сырых данных. *HBase* легко интегрируется с *Hadoop* для выполнения пакетной обработки больших данных. *HBase* обеспечивает высокопроизводительное хранение ключ–значение, что идеально подходит для хранения сырых данных о транспортных средствах. Эти данные включают статус транспортного средства, траектории движения, данные о действиях водителя и данные о заказах пользователей. В процессе принятия технических решений мы провели эксперименты по управлению сырыми данными о поездках с использованием СУБД, таких как *MySQL* и *PostgreSQL*.

Обзор существующих работ и технологий в области IoV. Работы в области *IoV* становятся всё более важными как в промышленности, так и в академических кругах. *IoV* является интегрированной технологией, включающей обработку больших данных, распределенные системы, управление большими данными, беспроводную связь, технологию конструирования транспортных средств, анализ человеческого поведения и другие аспекты. В облаке особенно критически важны обработка и управление большими данными.

Были предложены различные высокопроизводительные платформы для обработки больших данных, такие как *Storm*, *Spark*, *Samza*, *S4*, *Flume* и *Flink*. Однако по–прежнему сложно плавно интегрировать различные технологии для разработки систем для сложных сценариев в *IoV*.

Компании, предоставляющие услуги частного автотранспорта, такие как *Uber* и *Lyft*, разрабатывают сложные системы для обработки данных, интегрируя онлайн– и офлайн–обработку и управление большими данными для создания экосистемы для интеллектуальных транспортных средств.

Дизайн системы для потоковой обработки является сложной задачей исследования в *IoV*. На *Google* разработана модель потоковых данных для обработки неограниченных, неупорядоченных потоков данных. *Twitter* использует *Storm* для обработки потоковых данных. Компания *Amazon* предлагает *Kinesis* в качестве платформы для обработки больших данных. *Facebook* разрабатывает систему обработки данных в реальном времени. Предложены также различные системы для обработки и анализа больших пространственных данных, такие как *TrafficDB* и *Hadoop–GIS*.

Заключение. В данной статье описан опыт и решение проблем при создании *CarStream* – промышленной системы обработки больших данных для *IoV* и опыт построения нескольких приложений на основе этой системы для услуг автомобильного обслуживания с водителем. *CarStream* обеспечивает высокую надежность для сервисов, критичных для безопасности в *IoV*, включая трехуровневую систему мониторинга, охватывающую все слои от прикладного уровня до инфраструктурного. *CarStream* дополнительно использует кэширование в памяти и потоковую обработку для решения проблем обработки в реальном времени, обработки большого объема данных, низкого качества данных и низкой плотности информации. *CarStream* управляет большим объемом данных о поездках с помощью гетерогенной системы хранения данных. На

данный момент система может обрабатывать десятки тысяч автомобилей. В будущей работе планируется развивать *CarStream* в систему с архитектурой микросервисов для лучшего обслуживания системы и разработки новых приложений при увеличении числа автомобилей.

Список литературы

[1] R. Ranjan. Streaming big data processing in datacenter clouds. IEEE Cloud Computing, 1(1):78–83, 2014.

[2] M. Stonebraker, U. C¸ etintemel, and S. Zdonik. The 8 requirements of real–time stream processing. ACM SIGMOD Record, 34(4):42–47, 2005.

[3] M. Tang, Y. Yu, Q. M. Malluhi, M. Ouzzani, and W. G. Aref. LocationSpark: A distributed in–memory data management system for big spatial data. In Proc. VLDB Endow., 9(13):1565–1568, 2016.

Авторский вклад

Мигалевич Сергей Александрович – руководство исследованием потенциала современных информационных технологий в автомобильной промышленности.

Марков Алексей Николаевич – рассмотрение возможности совместного использования различных информационных технологий, технологий больших данных в автомобильной промышленности.

Ершов Денис Геннадьевич – исследование промышленной системы обработки больших данных для услуг аренды автомобилей с водителем.

INTEGRATION OF BIG DATA PROCESSING IN THE AUTOMOTIVE INDUSTRY

S.A. Migalevich
Master of Technical Sciences, Head of the Center for Informatization and Innovative Developments

A.N. Markov
Master of Technical Sciences, Deputy Head of the Center for Informatization and Innovative Developments

D.G. Ershov
Student of the Belarusian State University of Informatics and Radioelectronics

Abstract. As the Internet–of–Vehicles (IoV) technology becomes an in creasingly important trend for future transportation, de signing large–scale IoV systems has become a critical task that aims to process big data uploaded by fleet vehicles and to provide data–driven services. The IoV data, espe cially high–frequency vehicle statuses (e.g., location, engine parameters), are characterized as large volume with a low density of value and low data quality. Such characteristics pose challenges for developing real–time applications based on such data. In this paper, we address the challenges in de signing a scalable IoV system by describing CarStream, an industrial system of big data processing for chauffeured car services. Connected with over 30,000 vehicles, CarStream collects and processes multiple types of driving data includ ing vehicle status, driver activity, and passenger–trip infor mation. Multiple services are provided based on the col lected data. CarStream has been deployed and maintained for three years in industrial usage, collecting over 40 ter abytes of driving data. This paper shares our experiences on designing CarStream based on large–scale driving–data streams, and the lessons learned from the process of address ing the challenges in designing and maintaining CarStream.

Keywords: Big Data, knowledge management, automotive industry, Generic awareness.