**П.К. Куценогий,** ***кандидат физико-математических наук***
**Т.А. Лужных, В.С. Риксен**
*Сибирский федеральный научный центр агробиотехнологий РАН*
*РФ, 633501, Новосибирская область, п. Краснообск, ул. Центральная, Президиум*
E-mail: peter@kutsenogiy.ru

УДК 631.153 DOI: 10.30850/vrsn/2020/6/10-13

# ОПТИМИЗАЦИЯ СТАНДАРТОВ СБОРА ДАННЫХ О СЕЛЬСКОХОЗЯЙСТВЕННОЙ ДЕЯТЕЛЬНОСТИ В АНАЛИТИЧЕСКИХ ЦЕЛЯХ

*В статье рассмотрены важность и ценность использования больших данных (Big data) в сфере сельского хозяйства. Технологии «больших данных» предполагают обработку огромного объема разнообразных структурированных и неструктурированных данных, а также использование различных инструментов, подходов, методов их обработки, позволяющих анализировать информацию, необходимую для решения конкретных целей и задач. Извлечение информации из таких массивов данных и дальнейший ее интеллектуальный анализ открывают возможность сельхозпроизводителям улучшить качество принимаемых решений и определить своевременные, более эффективные методы ведения сельского хозяйства. Представлены разработанные шаблоны баз данных, максимально полно учитывающих особенности собираемой и анализируемой информации в интересах сельхозпроизводителей и отраслевых экспертов. В работе была поставлена задача – создать гибкую структуру, которую можно оперативно дополнить новой значимой информацией. Проведено тестовое наполнение разработанных шаблонов первичной информацией для проверки работоспособности создаваемой структуры базы данных. Часто статистика ведется не с целью получения объективной информации, а с некими «политическими» или рыночными манипуляциями. Дополнительно возникает проблема «прерывности» данных, собираемых в период очередных кампаний, которые через некоторое время сходят на нет. Использование стандартов для обеспечения взаимодействия между системами – ключевое требование эффективной интеграции информации. Учитывая опыт заполнения создаваемой базы данных, а также тестируемых возможностей использования, сформулированы требования к стандартам сбора данных сельхозпроизводителей.*
**Ключевые слова:** *большие данные (Big data), машинное обучение, база данных, сбор данных, стандарты.*

**P.K. Kutsenogiy,** ***PhD in Physico-mathematical sciences***
**T.A. Luzhnyh, V.S. Riksen**
*Siberian Federal Scientific Center for Agrobiotechnology RAS*
*RF, 633501, Novosibirskaya oblast', p. Krasnoobsk, ul. Central'naya, Prezidium*
Email: peter@kutsenogiy.ru

# OPTIMIZING STANDARDS FOR COLLECTING AGRICULTURAL DATA FOR ANALYTICAL PURPOSES

*The article discusses the importance of using big data in agriculture. Technologies of "big" imply the processing of a huge amount of data of various structured and unstructured data, as well as the use of various tools, methods of their processing, which allowing to analyze information, special solutions for specific goals and objectives. Extracting data from such data sets and improving its intellectual analysis, open up the opportunity for agricultural producers to improve the quality of decisions made and timely, more efficient methods of farming. The developed database templates are presented, they which fully take into account the features of the collected and analyzed information in the interests of agricultural producers and industry experts. The task was set in the work to create a flexible structure that can be quickly supplemented with new relevant information. A test filling of the developed templates with primary information to check the performance of the created database structure was carried out. Often statistics is undertaken not for the purpose for obtaining objective information, but for some kind of "political" or market manipulation. Additionally, there is a problem of "discontinuity" of data collected during the period of another campaigns, which after a while they're gone. Using standards to ensure interoperability between systems is a key requirement for effective information integration. Taking into account the experience of filling in the creating database as well as the tested possibilities of using, the requirements for the data collection standards of agricultural producers were formulated.*
**Ключевые слова:** *big data, machine learning, database, data collection, standards.*

Технологии «больших данных» (Big data, BD) предполагают обработку огромного объема разнообразных структурированных и неструктурированных данных, а также использование различных инструментов, подходов, методов их обработки, позволяющих анализировать информацию для решения конкретных целей и задач. [4, 10] Извлечение информации из таких массивов данных и дальнейший ее интеллектуальный анализ открывают возможность сельхозпроизводителям улучшить качество принимаемых решений и определить своевременные, более эффективные методы ведения сельского хозяйства. [9]

Необходимо учесть важность полноты и достоверности исходных данных для методов анализа Big data. Например, в процессе «машинного обучения без учителя» сами данные могут выступать источником алгоритмов, генерирующих результат интерпретации этих же данных. В этих условиях становится понятно, что ошибки будут автоматически вести к систематическим проблемам в их интерпретации. Следует отметить традиционные проблемы сбора сельскохозяйственных данных. Часто статистика ведется не с целью получения объективной информации, а с некими «политическим» или рыночными манипуляциями. Дополнительно

возникает проблема «прерывности», собираемых в период очередных кампаний данных, которые через некоторое время сходят на нет. [3]

Использование стандартов для обеспечения взаимодействия между системами – ключевое требование эффективной интеграции информации.

Основная трудность в достижении совместимости между несколькими хранилищами BD заключается в различиях метаданных, используемых в одном хранилище, относительно других. Без стандартов для этих метаданных интеграция данных, генерируемых в проектах BD, будет еще более сложной задачей. [5]

Интеграция BD осуществляется посредством Extract, Transform, Load (ETL) – одним из основных процессов в управлении хранилищами данных, который включает в себя три важные функции (рис. 1), необходимые для получения данных из одной среды и помещения их в другую. [8]

Первый шаг – извлечение. На этом этапе данные из исходных систем передаются в область подготовки, которые могут быть в разных форматах, таких как реляционные базы (БД). Важно эти данные сохранить в промежуточной области, а не в хранилище, поскольку извлеченные данные имеют различные форматы и могут быть повреждены. Следовательно, загрузка непосредственно в хранилище данных может повредить его, и откат будет намного сложнее. Поэтому это один из важнейших этапов процесса ETL.

Второй шаг – преобразование. К извлеченным данным применяется набор правил или функций для преобразования их в единый стандартный формат; последний – загрузка. Преобразованные данные загружаются в хранилище. Чаще всего они обновляются путем загрузки, а иногда через более длительные, но регулярные интервалы. Скорость и период загрузки зависят исключительно от требований и варьируются от системы к системе. [13]

Для интеграции BD используют инструменты: MapReduce, Spak, Powercenter, SQL Server Integration Services и другие. Взаимодействие локальных баз данных будет осуществляться посредством запроса SQL. [8]

Международной организацией по стандартизации (ISO) и Международной электротехнической комиссией (IEC) в сфере BD разрабатываются стандарты для управления данными внутри и между локальными и распределенными средами информационных систем. Один из таких стандартов – ISO / IEC TR 10032: 2003. Он имеет следующую область применения [7]:

– эталонные модели и структуры для координации существующих и новых стандартов;

– определение областей данных, типов данных и структур данных и связанных с ними семантик;


**Рис. 1. Процесс ETL.**

– языки, сервисы и протоколы для постоянного хранения, одновременного доступа и обновления, обмена данными;

– методы, языки, сервисы и протоколы для структурирования, организации и регистрации метаданных и других информационных ресурсов, связанных с совместным использованием и совместимостью, включая электронную торговлю.

При разработке шаблонов БД мы следовали требованиям стандарта:

1. Разрабатываемая нами БД хранится и обрабатывается в вычислительной системе.

2. Данные логически структурированы (систематизированы) с целью их эффективного поиска и обработки. Структурированность подразумевает явное выделение составных частей, связей между ними, а также типизацию элементов и связей, при которой с типом связи соотносится определенная семантика и допустимые операции.

3. БД включает схему, или метаданные, описывающие логическую структуру БД в формальном виде (некоторая метамодель).

В соответствии со стандартом, схема включает описания содержания, структуры и ограничений целостности, используемые для создания и поддержки БД. В базе находится набор постоянных данных, определенных с помощью схемы. Система управления использует определения данных в схеме для обеспечения доступа и управления доступом в БД. [8]

Для стандартов применяют методы оптимизации параметров объектов стандартизации (ПОС): установление значений, количественно характеризующих свойства объектов стандартизации (параметры или данные), при которых достигается максимально возможная, в определенных условиях, эффективность использования (максимальный эффект на единицу затрат). Значения ПОС, которым соответствует максимально возможная эффективность, называют оптимальными; аналогично называют и уровень требований стандартов.

При сборе данных о сельскохозяйственной деятельности необходимо учитывать ее специфику – непрерывность и повторяемость, то есть годовой сельскохозяйственный цикл. Кроме того, следует учитывать особенности работы хозяйств, сельхозпроизводителей, как источников данных, а также планируемые методы анализа. В нашем случае, мы ориентируемся на возможность анализа BD современными вычислительными методами.

*Описание создаваемых шаблонов базы данных*

Для построения шаблона БД использован Microsoft Office Access. Разнообразный спектр функций продукта включает связь с внешними таблицами и БД.

Шаблон БД сельскохозяйственного производства разработан в аналитических целях с учетом использования технологий BD и машинного обучения. При поступлении достаточного объема данных в соответствии с разработанным шаблоном, будут применены алгоритмы машинного обучения градиентного бустинга и бэггинга, которые решают задачи классификации, регрессии и кластеризации с помощью методов построения моделей XGBClassifier и Random Forest Classifier. [6, 11]

В создаваемой БД систематизируется и концентрируется информация о выращивании сельско-

хозяйственных культур в разных почвенно-климатических зонах России. Разработанные шаблоны заполнили данными реального сельскохозяйственного предприятия для определения практических требований к стандартам подобных данных. При взаимодействии с другими информационными системами и повторном использовании метаданных важное значение имеет стандартизация средств их представления. В настоящее время созданы многочисленные стандарты описания метаданных вертикальной и горизонтальной сферы. Активно используют стандарты платформы XML (ISO 20022-4), Дублинское ядро, дескриптивное множество языка SQL (ISO 9075). [2, 3]

БД имеет иерархически организованную реляционную структуру (совокупность таблиц «сущность – связь») с перекрестными ссылками, использующими уникальные ключи записей ID (рис. 2). Верхний уровень иерархии включает две таблицы сельхозпредприятий и метеостанций. Оболочка системы определяет ближайшую метеостанцию, затем погодные данные перекачиваются в БД и преобразуются в вид, представленный на рис. 3.

Полевой опыт – следующий уровень иерархии БД. Ссылка на таблицу несколько условна: если в расчете необходимо использовать данные с производственных посевов, то вместо имени полевого опыта следует записывать – «производственные посевы». В графу год начала опыта заносится дата начала наблюдений.

Таблица годовой ротации поля – следующая по уровню иерархии – соответствующие записи объединяют ссылки на год, культуру и поле.

Справочник культур содержит перечень сортов сельскохозяйственных культур, возделываемых на территории Российской Федерации, входящих в реестр селекционных достижений.

В справочник типов почв, входит информация о возможных в данной зоне типах почвы с соответствующим ID номером. Ее можно пополнять при расширении на другие регионы.

Перечень технологий (таблица), используемых в каждом варианте опыта, это фактически только ссылка на идентификатор технологии для определенного года в конкретном севообороте.

Все последующие таблицы включают расшифровку применяемой технологии. В них содержатся данные о севе, механической обработке почвы, внесении удобрений и режиме их применения. Все таблицы построены однотипно. Каждая операция сопровождается описанием технологии, которая может применяться неоднократно.

Таблица урожаев содержит данные о сроках уборки и валовом сборе, для трав включаются показатели по укосам.

Вся содержательная информация, хранящаяся в БД, может быть разделена на два основных класса – постоянная (условно) и оперативная (наблюдения и измерения). Данные первого класса вводятся в БД либо однократно, либо с периодичностью в несколько лет. К таким показателям относятся, например, агрохимические характеристики почвы, измеряемые один раз в пять лет. Оперативные данные вводят в БД по мере их поступления в каждом сезоне вегетации.


**Рис. 2. Упрощенная концептуальная схема БД.**

В процессе заполнения описанной БД были выявлены обязательные требования к структуре, полноте и виду данных для того, чтобы их можно было использовать для обработки методами машинного обучения. Пробелы в данных не дают возможности сформировать необходимое количество «обучающих примеров», а структура формирует данные, относящиеся к задаче, которую мы пытаемся решить.

Информация зачастую формируется из различных источников. Среди них могут оказаться нерелевантные или ненужные значения, которые потребуется удалить, а каких-то может не хватать, и их необходимо добавить. От правильной подготовки базы данных зависит и пригодность к использованию, и достоверность результатов.

*Сформулируем требования к стандартам сбора данных.*

Для каждой единицы данных (ячейка) необходима привязка к месту (географические координаты) и ко времени, где и когда осуществлялся сбор данных.

Последовательные временные ряды данных должны быть без пробелов и пропусков.

В дальнейшей работе необходимо выявить возможные избыточные требования к стандартам сбора данных, которые не увеличивают объем информации, содержащийся в данных, а усложняют их получение и накопление. Следует учесть, что

Погодные данные

| Имя поля | Тип данных | |
|---|---|---|
| Код | Счетчик | |
| Код метеостанции | Числовой | |
| Синоптический индекс | Числовой | |
| Дата | Дата и время | б/р |
| Температура средняя | Числовой | град. С |
| Минимальная температура во | Числовой | град. С |
| Температура точки росы | Короткий текст | град. С |
| Максимальная температура во | Числовой | град. С |
| Относительная влажность воз | Числовой | % |
| Влажность воздуха мин | Числовой | % |
| Атмосферные осадки | Числовой | мм |
| Средняя скорость ветра | Числовой | м/с |
| Порыв | Числовой | м/с |
| Минимальная видимость | Числовой | м |
| Направление ветра | Короткий текст | б/р |
| Атмосферное давление Р ср | Числовой | гПа |
| Атмосферное давление Р мин | Числовой | гПа |
| Атмосферное давление Р макс | Числовой | гПа |

**Рис. 3. Таблица погодных данных.**

важными источниками в перспективе будут сельскохозяйственные производители – пользователи экспертного ресурса. Хозяйства могут быть совершенно разного экономического масштаба: от мелких фермерских до вертикально интегрированных агрохолдингов. Понятно, что в кадровом составе небольших предприятий может и не быть IT-специалистов, ответственных за внесение данных в сложных аналитических системах. Поэтому, возникает еще два требования к стандартам сбора:

– данные могут вноситься в любом цифровом формате, понятном для пользователя и совместимом с наиболее распространенными цифровыми стандартами, позволяющими экспортировать данные для дальнейшей обработки;

– при формировании списка параметров, подлежащих регулярному мониторингу, следует избегать данных, для сбора которых необходимы приобретение и эксплуатация дорогостоящего и сложного оборудования.

**Заключение.** На данном этапе мы разработали структуру БД. Информация, вносимая пользователем, будет способствовать наиболее полному и объективному анализу в интеллектуальной системе. Предлагаемая структура БД позволяет встроить ее в комплексный вычислительный процесс, используя стандартные модули преобразования входных и выходных данных, широко применяемые стандарты, быстро переформатировать данные под нужную задачу. БД также будет наполняться архивной и актуальной информацией, которая может быть использована в ином формате. Сформулированы основные требования, предъявляемые к собираемым данным, а также отмечены возможные, которых следует избегать в целях оптимизации данных стандартов. В дальнейшем планируется расширить продукт с дружеским пользовательским интерфейсом, цель которого – сбор статистических данных агропредприятия. Второй планируемый модуль – модуль загрузки данных со сторонних ресурсов, таких как серверы, содержащих данные статистики, погодные условия и др.

**СПИСОК ИСТОЧНИКОВ**


1. Когаловский, М.Р. Метаданные в компьютерных системах / М.Р. Когаловский // Наука «Интерпериодика». – 2013. – Т. 39 – № 4. – С. 28–46.
2. Нафикова, А.Р. Практическое руководство по Microsoft SQL Server / Учебное пособие // А.Р. Нафикова – Стерлитамак: Стерлитамакский филиал БашГУ. – 2014. – 124 с.
3. Чубукова, И.А. Data Mining: учебное пособие /2-е изд., испр. //И.А. Чубукова – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний. – 2008. – 382 с.
4. Coble, K. Big data in agriculture: a challenge for the future / K. Coble, A. Misra, S. Ferrell, T. Griffin // Applied Economic Perspectives and Politics. – 2018. – P. 79–96.
5. Hammer, C. Big Data: Potential, Challenges, and Statistical Implications / C. Hammer, D. Kostroch //CUSSION NOTE. – 2017. – P. 11–15.
6. Hastie, T. Chapter 15. Random Forests / T. Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer-Verlag, 2009. – 746 p.
7. ISO/IEC 2003 Information technology – Reference Model of Data Management // [electronic resource]. 2003. https://www.iso.org/files/live/sites/isoorg/files/developing_stanards/docs/en/big_data_report-jtc1.pdf.
8. Jamack, P. Hive as a tool for ETL or ELT / P. Jamack // IBM. – 2014. – P. 4–8.
9. Park, H. Recent advancements in the Internet-of-Things related standards / H. Park, H. Kim, H. Joo, J. Song // A oneM2M perspective. ICT Express, vol. 2, no. 3. – 2016. – P. 126–129.
10. Ribarics, P. Big data and its impact on agriculture /P. Ribarics // Ecocycles 2 (1). – 2016. – P. 31–35.
11. Tseng, G. Gradient Boosting and XGBoost / G. Tseng // [electronic resource]. 2018. https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5.
12. Vasileios, T. Frequent patterns in ETL workflows: an empirical approach / T. Vasileios // Data and knowledge engineering. – 2017. – P. 1–16.


**LIST OF SOURCES**


1. Kogalovskij, M.R. Metadannye v komp'yuternyh sistemah / M.R. Kogalovskij // Nauka «Interperiodika». – 2013. – T. 39 – № 4. – S. 28–46.
2. Nafikova, A.R. Prakticheskoe rukovodstvo po Microsoft SQL Server / Uchebnoe posobie // A.R. Nafikova – Sterlitamak: Sterlitamakskij filial BashGU. – 2014. – 124 s.
3. Chubukova, I.A. Data Mining: uchebnoe posobie /2-e izd., ispr. //I.A. CHubukova – M.: Internet-Universitet Informacionnyh Tekhnologij; BINOM. Laboratoriya znanij. – 2008. – 382 s.
4. Coble, K. Big data in agriculture: a challenge for the future / K. Coble, A. Misra, S. Ferrell, T. Griffin // Applied Economic Perspectives and Politics. – 2018. – R. 79–96.
5. Hammer, C. Big Data: Potential, Challenges, and Statistical Implications / S. Hammer, D. Kostroch //CUSSION NOTE. – 2017. – R. 11–15.
6. Hastie, T. Chapter 15. Random Forests / T. Hastie, R. Tibshirani, J. Friedman // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer-Verlag, 2009. – 746 p.
7. ISO/IEC 2003 Information technology – Reference Model of Data Management // [electronic resource]. 2003. https://www.iso.org/files/live/sites/isoorg/files/developing_stanards/docs/en/big_data_report-jtc1.pdf.
8. Jamack, P. Hive as a tool for ETL or ELT / R. Jamack // IBM. – 2014. – R. 4–8.
9. Park, H. Recent advancements in the Internet-of-Things related standards / N. Park, H. Kim, H. Joo, J. Song // A oneM2M perspective. ICT Express, vol. 2, no. 3. – 2016. – R. 126–129.
10. Ribarics, P. Big data and its impact on agriculture /R. Ribarics // Ecocycles 2 (1). – 2016. – R. 31–35.
11. Tseng, G. Gradient Boosting and XGBoost / G. Tseng // [electronic resource]. 2018. https://medium.com/@gabrieltseng/gradient-boosting-and-xgboost-c306c1bcfaf5.
12. Vasileios, T. Frequent patterns in ETL workflows: an empirical approach / T. Vasileios // Data and knowledge engineering. – 2017. – R. 1–16.