УДК 631.58: 004.65

АНАЛИЗ И ВНЕДРЕНИЕ БОЛЬШИХ ДАННЫХ В АГРОНОМИИ

Эмилия Николаевна Аникьева
старший преподаватель
korol_0909@mail.ru
Семен Антонович Мацко
студент
matsko.sema13@mail.ru
Мичуринский государственный аграрный университет
г. Мичуринск, Россия

Аннотация. Рассмотрены характерные особенности больших данных применительно к организации, сбору, хранению и доступу к информации в агрономии. На частном примере поиска данных в облачной модификации Google Big Table показана организация большой таблицы данных по вредителям, и болезням культур, выращиваемых преимущественно в центрально черноземном регионе.

Ключевые слова: большие данные, поиск, базы данных NoSQL, облачные технологии, агрономия.

Остановимся на первичных признаках и свойствах больших данных (Big Data), привлекательных в отношении методов их анализа при производстве продукции сельского хозяйства. Прежде, чем приступить к анализу, определим предмет наших исследований, чтобы в дальнейшем понимать, в какой степени мы излагаем именно то, что известно всем участвующими в этом сторонам. К настоящему времени нет общепринятого определения больших данных (БД), однако мы можем сформулировать некоторое правило, по которому можно узнать, о чем идет речь. Большими данными называют совокупность неструктурированных или частично структурированных данных большого объема, обработка, хранение и анализ которых невозможен обычными методами. Следовательно, мы можем немедленно выделить основные свойства БД.

1. Объем данных. Нет конкретного значения, после которого можно назвать объем данных большим, очень большим или гигантским. По некоторым подсчетам к большим данным можно отнести объемы, нарастающие со скоростью порядка 150 Гбт в сутки.

2. Скорость поступления данных. Объем данных постоянно меняется, можно уподобить БД потоку, протекающему через порты всех ваших задействованных в работе компьютеров, причем формат данных разнообразный. В качестве примера можно привести сбор и обработку данных поступающих с метеостанции непрерывно в течение месяца о температуре, давлении, скорости и направлении ветра.

3. Разнообразие данных. Имеется в виду разнообразие типов и форм представляемых данных, которые могут быть одновременно структурированными, частично структурированными и неструктурированными. Например, данные метеостанции об изменении температуры, влажности, давления, силы и направлении ветра, считываемые датчиками, фотографии посевных площадей, пересылаемые дронами, пространственно–временное сканирование местности спутниками, сводки отчетности о расходах на полив и сюда присовокупите также измерение влажности почвы, оценку индекса

благоприятного состава для высева определенной культуры. И это только часть информации, которая нужна для принятия правильных решений в эксплуатации посевных площадей.

4. Достоверность данных. Насколько полученные данные отражают реальное состояние процесса. Это свойство относится не только к самим данным, но также и к результатам их анализа.

5. Варьируемость или изменчивость данных. Данные могут значительно изменяться в зависимости от сезона, социальных явлений, политической повестки, и прочих факторов. Нельзя исключать ситуацию с внезапным изменением данных, так называемым выбросом или черным лебедем, - событием маловероятным, но предвестником существенного изменения дальнейшего хода процесса.

6. Ценность или значимость данных. Свойство, оказывающее наиболее существенное влияние на характеристики объекта изучения. Анализ таких данных позволяет построить модель прогнозируемого поведения объекта в течение достаточно долгого времени. Например, данные об изменении климата (температура, скорость ветров, прогнозы наводнений, выдувание плодородной почвы в засушливых районах, избыточное выпадение дождей, недостаток снежного покрова) полученные из различных по форме источников позволяют построить модели кратковременного и долговременного прогноза урожайности и изменении состава высеваемых культур, наиболее пригодных для адаптации в изменяющихся условиях локального географического положения агрохозяйства.

7. Расширяемость и масштабируемость. Новые поля в каждой группе данных могут быть легко добавлены или изменены, а размер системы хранения элемента или группы данных может быстро расширяться. В этом контексте выделяют также горизонтальную масштабируемость - возможность обработки данных одновременно на многих серверах без потери производительности.

В качестве технологий обработки больших данных к настоящему времени разработаны и применяются пять:

1. Нереляционные базы данных NoSQL (Not only SQL – не только SQL);

2. MapReduce – двух этапная обработка данных, разработанная компанией Google и использующая операции функционального программирования. На первом этапе (Map) главный компьютер (master node) получает данные решаемой задачи, разбивает их на отдельные задачи более низкого ранга и распределяет их между рабочими компьютерами (worker node) для промежуточной обработки. На втором этапе (Reduce – сведение, свертка) главный компьютер получает результаты решения разделенных задач низшего ранга от рабочих компьютеров и на этой основе формирует результат решения задачи. Преимущество такой технологии – возможность параллельного решения множества задач второго уровня на большом количестве серверов.

3. Hadoop – свободно распространяемый набор утилит, библиотек и программного обеспечения, предназначенный для реализации контекстного поиска на вебсайтах и социальных сетях, характеризующихся высокой степенью загруженности. Крос-платформенный модуль разработан на языке Java (аппаратная платформа – Java Virtual Machine) и структурно реализован на идее и модели MapReduce. Является проектом Apache Software Foundation и к настоящему времени характеризуется как основная технология обработки больших данных.

4. R – язык программирования и программное обеспечение для статистического анализа и обработки данных.

5. Business Intelligence – совокупность компьютерных методов и инструментов для анализа внешних рыночных показателей и внутренних финансовых, производственных показателей которые позволяют сформировать полную картину бизнес- решений, включая цели, направления и перспективы ведения бизнеса в данном секторе рынка.

Мы остановимся более подробно на технологии NoSQL имея в виду обработку больших данных в агропромышленном секторе.

Сбор данных и их обработка в аграрной промышленности можно показать в виде примерной схемы на рисунке.

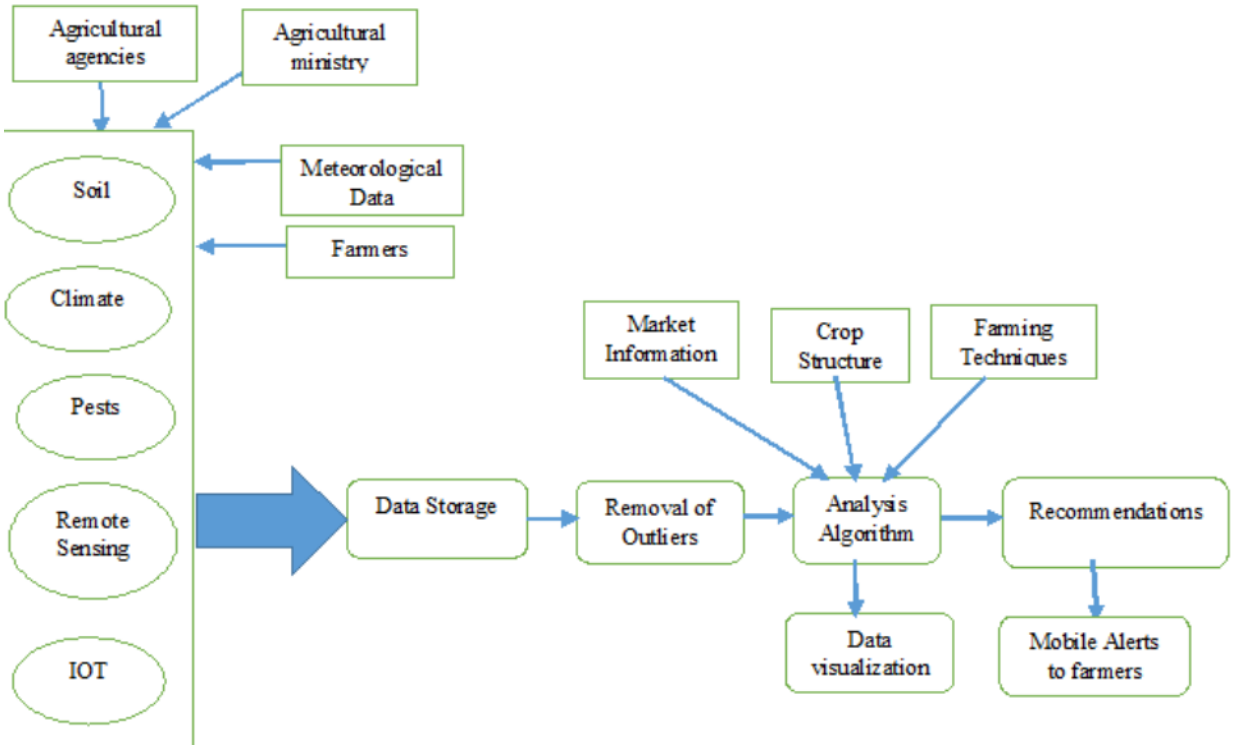

Рисунок 1 – Возможная картина сбора и обработки данных в сельском хозяйстве [1].

Данные о составе почвы, климате, вредителях, дистанционном зондировании площадей, получаемых со спутников и дронов, а также интернета вещей направляются в хранилище данных. Состав данных пополняется из различных источников: сельскохозяйственных агентств, министерства сельского хозяйства, метеостанций, фермерских хозяйств, датчиков температуры почвы, влажности, насыщенности гербицидами, видами вредителей и так далее. Очевидно, собранные данные не являются структурированными и имеют различную форму представления. Далее происходит предварительная обработка данных, устраняющая аномальные значения. После предварительной очистки от аномальных выбросов, данные поступают на анализ различными средствами анализа. Однако результат анализа таких данных в задаче об эффективности сбора урожая и сохранения продуктивности агросектора в значительной степени зависит от состояния рынка сельхозпродукции, структуры культур сбора и технического состояния фермерских хозяйств. Поэтому в алгоритме анализа больших данных должны

быть представлены также и такие данные. Результаты анализа могут быть выданы в виде рекомендаций агрохозяйствам в виде мобильных оповещений или представлены в визуальном виде на носителях информации с последующей распечаткой или храниться в базе данных хозяйства.

Обработка неструктурированных или частично структурированных данных производится, например, с помощью распределенной базы данных NoSQL.

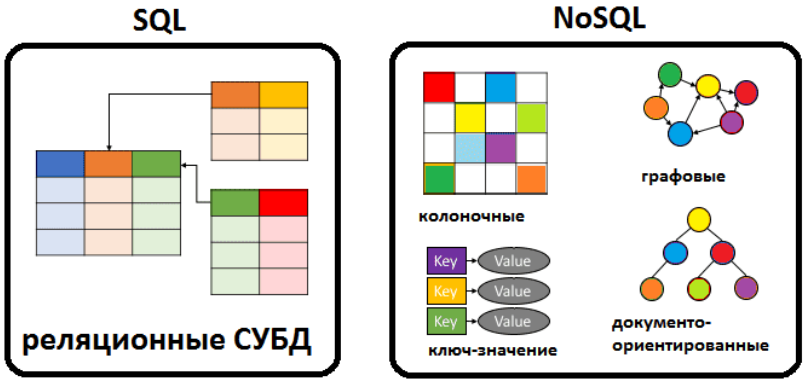

Рисунок 2 – Сравнение реляционной и не реляционной базы данных.

Наиболее общие типы баз данных NoSQL следующие:

•Колоночно-ориентированные – в них данные упорядочиваются по столбцам, а не по строкам. Это позволяет минимизировать объем информации и делает быстрым запрос, доступ и извлечение данных. Например, вам надо рассчитать общий объем поставленной на рынок продукции по региону. В этом случае вам нужен доступ только к двум колонкам – «Продажи» и «Регион».

•Ключ – значение – базы, организующие данные в виде уникальных ключей и связанной с ними информацией в кэше – значением.

•Документо-ориентированные базы данных, хранящие информацию в различных форматах документов;

•Графовые базы данных предназначены для хранения, управления и запросов информации в виде графов, состоящих из узлов и ребер.

Покажем на примере Google BigTable как может строиться колоночная база данных в аграрном комплексе.

Составим таблицу болезней и вредителей для некоторых культивируемых в ЦЧР растений: озимая пшеница, картофель, подсолнечник, сахарная свекла, просо, гречиха, кукуруза.

Таблица 1

Болезни и вредители для некоторых культивируемых в ЦЧР растений.

Культура (Crop)	Болезни (Deseases)			Вредители (Pests)	Обра-ботка	Выход %
	Гриб ковые	Бакте ри альные	Вирус ные			
Озимая пшеница	Снежная плесень, гниль, Септориоз Мучнистая роса, Ржавчина, головня			Гессенская муха, Опомиза, озимая совка Клоп, злаковые тли, хлебная жужжелица, грызуны		74
Картофель	Альтернариоз, антракноз, бурая гниль, кольцевая гниль, парша, рак картофеля			Колорадский жук, проволочник, медведка		78
Подсолнечник	Ржавчина, фомоз, альтернариоз, септориоз			Подсолнечная моль, медведка, проволочник, личинки жуков чернотелок, гусеницы, долгоносики		82
Сахарная свекла	Церкоспороз, мучнистая роса, альтернариоз, фоиоз, фузариоз, рамуляриоз			Проволочники (дичинки жуков щелкунов), личинки жуков чернотелок, гусеницы, свекловичная корневая тля, свекловичная нематода		69
Просо	Головня, гельминтоспориоз, Бактериальная пятнистость, некротический меланоз			Кукурузная тля, озимая совка, просяной комарик		88
Гречиха	Аскохитоз Белая пятнистость, гельминтоспориозная корневая гниль, фитофтороз, ложная мучнистая роса, спорынья, церкоспороз			Гречишная блоха, гречишный долгоносик, гречишный комарик, гречишная		81

		блошка, тля, пшеничная совка, проволочник, капустная совка, итальянский прус		
Кукуруза	Пузырчатая головня, корневая гниль, южный гельминтоспориоз, угольная гниль, бактериоз початков, вирус полосатой мозаики	Кукурузная тля, кукурузный мотылек, луговой мотылек, совки, цикадки кукурузные		91

Пример поиска строки таблицы из библиотеки клиента (Таблица Д) в облачной версии Goofle Big Table с помощью библиотеки кодов [2]:

```
#include "google/cloud/bigtable/table.D"

int main(int argc, char* argv[]) try {
  if (argc != 4) {
    std::string const cmd = argv[0];
    auto last_slash = std::string(cmd).find_last_of('/');
    std::cerr << "Usage: " << cmd.substr(last_slash + 1)
              << " <project_id> <instance_id> <table_id>\n";
    return 1;
  }

  std::string const project_id = argv[1];
  std::string const instance_id = argv[2];
  std::string const table_id = argv[3];

  // Create a namespace alias to make the code easier to read.
  namespace cbt = ::google::cloud::bigtable;

  cbt::Table table(cbt::MakeDataConnection(),
                   cbt::TableResource(project_id, instance_id, table_id));

  std::string row_key = "озимая пшеница";
  std::string column_family = "diseases";

  std::cout << "Getting a single row by row key:" << std::flush;
  google::cloud::StatusOr<std::pair<bool, cbt::Row>> result =
      table.ReadRow(row_key, cbt::Filter::FamilyRegex(column_family));
  if (!result) throw std::move(result).status();
  if (!result->first) {
    std::cout << "Cannot find row " << row_key << " in the table: " << table_id
              << "\n";
    return 0;
  }
```

```
  cbt::Cell const& cell = result->second.cells().front();
  std::cout << cell.family_name() << ":" << cell.column_qualifier() << "    @ "
            << cell.timestamp().count() << "us\n"
            << "" << cell.value() << "" << "\n";

  return 0;
} catch (google::cloud::Status const& status) {
  std::cerr << "google::cloud::Status thrown: " << status << "\n";
  return 1;
}
```

Список литературы:

1. Madhuri J, Indiramma M. Role of Big Table in Agriculture. International Journal of Innovative Technology and Exploring Engineering (IJITEE). Volume-9 Issue-2, December 2019

2. Google Developers Codelabs — это обучающие практические занятия по программированию. – URL: https://codelabs.developers.google.com/

UDC 531.58: 004.65

ANALYSIS AND IMPLEMENTATION OF BIG DATA IN AGRONOMY

Emilia N. Anikyeva
senior lecturer
korol_0909@mail.ru
Semen An. Matsko
student
matsko.sema13@mail.ru
Michurinsk State Agrarian University
Michurinsk, Russia

Abstract. The characteristic features of big data are considered in relation to the organization, collection, storage and access to information in agronomy. A specific example of data search in the cloud modification of Google Big Table shows

the organization of a large table of data on pests and diseases of crops grown mainly in the Central Black Earth Region.

Keywords: big data, search, NoSQL databases, cloud technologies, agronomy.

Статья поступила в редакцию 10.05.2025; одобрена после рецензирования 20.06.2025; принята к публикации 30.06.2025.

The article was submitted 10.05.2025; approved after reviewing 20.06.2025; accepted for publication 30.06.2025.