ПОЧЕМУ ЭЛЕКТРОННЫЕ ТАБЛИЦЫ ОГРАНИЧИВАЮТ ВОЗМОЖНОСТИ АНАЛИЗА ДАННЫХ

МАЙКЛ РИССЕ (MICHAEL RISSE)
ПЕРЕВОД: ВЛАДИМИР РЕНТЮК

В статье дается ответ на вопрос, почему для очистки данных, визуализации, поиска, контекстуализации и моделирования необходимы соответствующие инструменты. Представлены возможности современных программных средств для анализа информации.

Компании, занятые в перерабатывающей отрасли, собирали свои производственные данные десятилетиями. Однако сегодня ситуация складывается таким образом, что с каждым новым шагом в развитии аппаратного и программного обеспечения в промышленности генерируется и концентрируется еще больше сведений, объединенных термином «большие данные». Они характеризуют условия технологических и производственных процессов, предоставляют информацию по цепочкам поставок и отражают ряд других не менее важных производственных характеристик.

В связи с все более увеличивающимся количеством данных компании стремятся преобразовать такие объемы в полезную информацию о текущем положении дел на предприятии и в стратегию его дальнейшего развития. Это делается для повышения надежности, безопасности и рентабельности технологических установок на заводах и предприятиях в целом. Но, как бы они ни старались, с ростом объемов данных проблемы только усиливаются и множатся.

В свою очередь четвертая промышленная революция, или «Индустрия 4.0», вызванная индустриальным «Интернетом вещей» (Industrial Internet of Things, IIoT), также разворачивается на основе передовых технологий компьютеризации, распространения сенсоров и беспроводных технологий, значительно расширяя типы данных и объемы для их хранения и анализа.

Исторически сложилось так, что компании, связанные с непрерывным производством, для организации данных, собранных в виде таблиц, использовали электронные варианты отображения. Первоначально такие таблицы предназначались для учета и финансов и обычно никогда не предназначались для больших объемов данных, тем более в виде временных рядов. Однако позволяли создавать программируемые формулы, а также проводить расчеты на нескольких листах и книгах (по терминологии Excel от компании Microsoft).

Поэтому инженеры и приняли электронные таблицы для анализа данных, что впоследствии привело к трудоемким и затратным по времени процессам. Кроме того, было крайне сложно использовать электронные таблицы, когда требовалось обмениваться результатами и кооперироваться с другими подразделениями даже в рамках одного предприятия, не говоря уже о внешнем обмене информацией с другими организациями. Поскольку компании накапливали все больше данных, то всеми силами пытались найти эффективные способы обмена ими, хотя бы внутри своих подразделений.

ПРОБЛЕМА ОБЪЕМОВ ДАННЫХ

Для преодоления этих проблем и барьеров служат передовые средства (в частности, программное обеспечение) анализа данных. Чтобы осмыслить эти достижения и открывающиеся при их использовании возможности, рассмотрим четыре ограничения электронных таблиц и то, как каждое из них можно преодолеть с помощью новых аналитических решений.

Современные системы производства и мониторинга технологических и производственных процессов, как уже было сказано, создают огромные объемы данных, которые в совокупности характеризуют условия процесса, текущие операции изготовления тех или иных продуктов и состояние оборудования. Причем данные, относящиеся к технологическим системам и системам управления, генерируются в различных формах. Общий подход состоит в том, чтобы собрать все родственные данные, связанные с конкретным производственным процессом, в электронную таблицу, а затем выполнить их анализ. Однако огромный объем собранной информации, причем из нескольких источников, быстро подрывает возможности для проведения такого анализа эффективно и однозначно.

Перед выполнением аналитики все поступившие сведения должны быть отсортированы и очищены, а количество точек данных в элек-

тронной таблице уменьшено. При этом сигналы от инструментальных средств переформатируются в данные, которые соответствовали бы парадигме столбец/строка электронной таблицы, как это показано на рис. 1. Заявленный предел для электронной таблицы Microsoft Excel составляет около 1 млн строк. Обычная частота выборки датчиков системы типового технологического процесса, например со съемом показателей один раз в 1 мин, соответствует полумиллиону строк в Excel в год. Если частота выборки составляет один раз за каждые 30 с или если пользователь хочет ознакомиться с информацией за два года, в таком случае просмотреть все данные в правильном разрешении уже невозможно.

Кроме того, файлы, расширяющие пределы емкости электронных таблиц, будут приводить к проблемам с производительностью. Слои в нескольких наборах данных и расчетов, одновременное открытие множества больших файлов, ссылки на другие приложения и макросы препятствуют удобной работе с электронными таблицами. Так что при использовании этих таблиц приходится идти на компромисс в отношении типа и выборки сегментов данных. Однако для инженера, контролирующего ход процесса, или ученого, занятого разработкой процесса, как правило, нет мелочей и требуются все перечисленные возможности по анализу данных.

ПРОБЛЕМА ИЗОЛЯЦИИ ДАННЫХ

Изоляция данных, хотя это обычно связано с ограничениями по объему, представляет собой отдельную проблему. Например, каждый раз, когда член команды обращается к данным процесса, он сначала загружает их в отдельный дублирующий файл. Это разовое извлечение «снимка» (назовем так отображение текущего состояния) процесса. Если данные изменяются или обновляются, запрос должен быть послан снова, а «снимок» переделан. Это может иметь те или иные, в том числе и крайне негативные, последствия для дальнейших расчетов, очистки данных и понимания происходящего. Кроме того, большие по объемам файлы трудно распространять по всей организации и синхронизировать, особенно если несколько пользователей просматривают одни и те же наборы данных из одних источников.

Учитывая возможности IIoT и облачных технологий, создание все более емких баз данных является постоянной и нарастающей тенденцией. Кроме того, не все данные, базы и пользователи находятся в одном месте. Еще больше усложняют задачу получения правильных сведений удаленные базы данных и пользователи.

Есть еще одна проблема. Когда соответствующие данные собраны в электронной таблице, как пользователям находить информацию, основанную на таком представлении информации? Инженеры больше всего интересуются тем, как данные по оборудованию и процессу ведут себя во времени и по отношению к другим системным элементам. Например, их интересует температура, давление, качество сырья и степень и эффективность его переработки, сроки выполнения тех или иных операций — все это имеет отношение к переработке сырья в готовый конечный продукт. Кроме того, необходимо следить за состоянием оборудования и вовремя принимать, желательно превентивные, меры по его техническому обслуживанию.

Как и в любом анализе, пользователь должен сначала определить точки процесса, представляющие наибольший интерес: оптимальные установившиеся условия, развитие вибраций критического оборудования, отключения, выбросы и другие параметры. И здесь именно время является критическим фактором. Инженеры, для того чтобы выявить тенденции и коренные причины тех или иных отклонений, анализируют данные, сгруппированные по сменам, неделям, месяцам или даже годам.

Для того чтобы сделать это в электронной таблице, пользователи, которые должны определить точки данных для рассмотрения, сортируют столбцы и строки. Эта сортировка/очистка выполняется систематически с помощью функций электронных таблиц, но 70% из 10 наиболее часто используемых функций списков Microsoft для Excel предназначено для обработки, а не для анализа данных во времени, из чего и достигается их ценность в рассматриваемом аспекте.

Манипулирование данными составляет 50–90% времени, затрачиваемого на создание приложений для работы с электронными таблицами, как это показано на рис. 2. Алгоритмы работы с электронными таблицами могут сортировать и разрезать данные, но подходы к манипуляциям с данными/вычислениями не прозрачны, и их может быть сложно запомнить и поделиться с коллегами.

Для примера возьмем ежемесячный отчет или ежеквартальную оценку выбросов, для этого нам необходимо запрашивать соответствующие данные, а любые элементы, требующие ручного ввода, должны воспроизводиться или автоматизироваться с помощью макро-

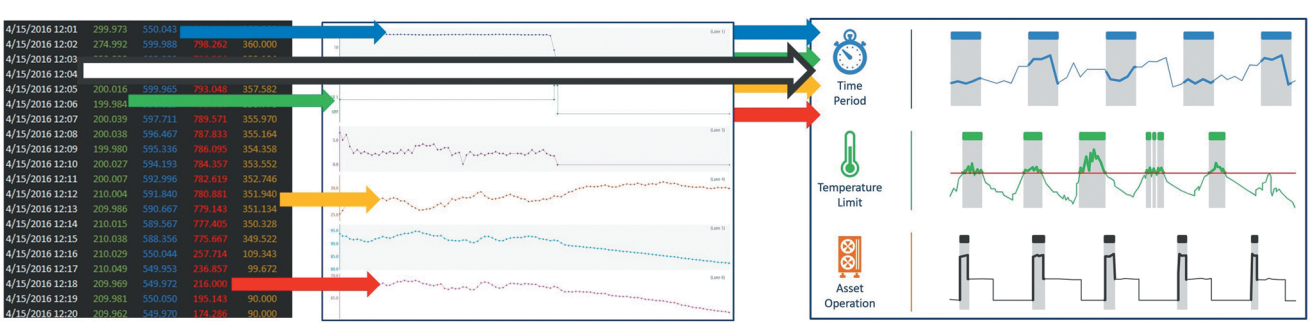


РИС. 1. ▼ Ключевым элементом при оценке данных технологического или производственного процесса является время. Для использования в электронных таблицах часто необходимо переформатировать показания датчиков, чтобы привести их в единую форму с информационными сигналами. Все рисунки предоставлены компанией Seeq Corp.

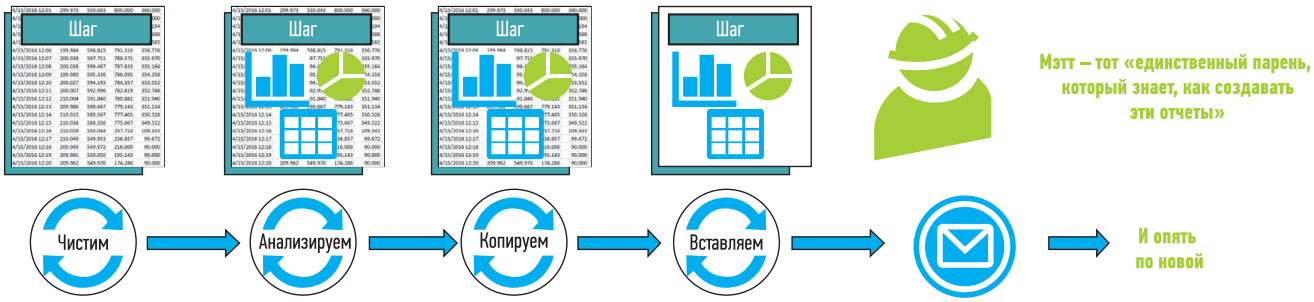


РИС. 2. ▲ Выделение, основанное на данных, полученных в результате анализа электронных таблиц. Последующее распространение этой информации является трудоемким и длительным процессом

РИС. 3. ▼ Группа управления энергопотреблением и группа усовершенствования технологического процесса большую часть своего времени потратили на подготовку анализа данных, а не на их непосредственный анализ. При этом только один человек мог понять и работать с таблицей для создания отчета

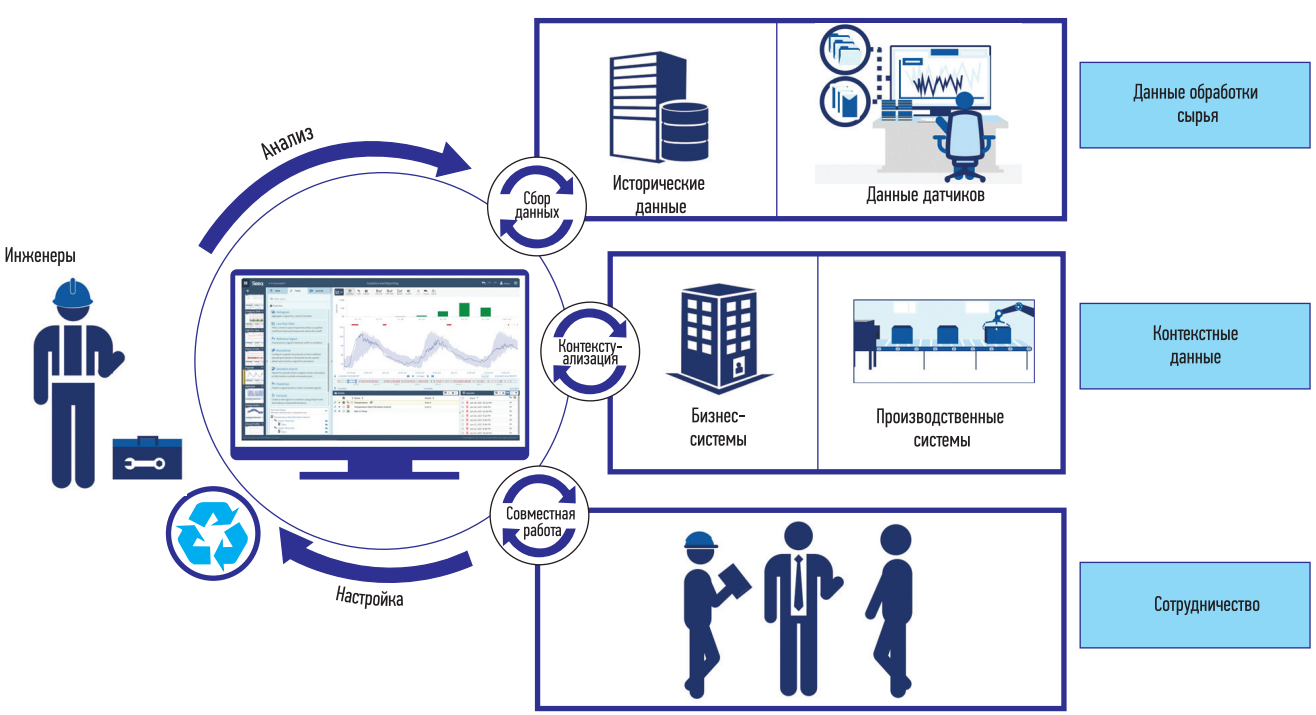
сов. Если анализ проводится нечасто или разными лицами, то для рассмотрения или повторного изучения манипуляций с данными, представленными в виде электронной таблицы, может потребоваться значительное время. И хотя некоторые предприятия имеют даже отдельную документацию для описания рабочих процессов, отсутствие прозрачности при разработке макросов затрудняет воспроизведение практически любого анализа.

ОГРАНИЧЕННЫЕ ВОЗМОЖНОСТИ СОТРУДНИЧЕСТВА И СЛОЖНАЯ ОТЧЕТНОСТЬ

Давайте зададим себе еще один вопрос. Как большие массивы данных сортируются и просеиваются, как разделяются и распространяются данные, управляемые данными? Что касается первой части вопроса, то нам доступны ограниченные возможности для извлечения информации из анализа электронных таблиц. Что же по второй части, непрозрачность вычислений затрудняет совместную работу и воспроизведение результатов. Да и по соображениям размера и простоты, как правило, общим является конечное изображение анализа — результат, а не сама таблица.

Кроме того, из-за изоляции данных и жестких ограничений на манипуляции ими работа, выполняемая с помощью электронных таблиц, должна быть централизованно доступна и тщательно поддерживаться. Это становится весьма проблематичным, когда результаты передаются в другую форму для их распространения.

Отчеты и совместное использование данных часто состоят из копирования и вставки или рабочего процесса в виде «вставка/ссылка/запрос/повторение запроса» (если произошел сбой в файле). Устранить эти и другие ограничения, характерные для электронной таблицы по самой ее природе, позволяет усовершенствованное аналитическое программное обеспечение. Его применение обеспечивает более быстрое и однозначное понимание процессов, как это будет показано в следующем примере.



ПРИМЕР АНАЛИЗА И ЭКОНОМИИ ЭНЕРГИИ ПРИ ПЕРЕРАБОТКЕ ЗЕРНОВЫХ КУЛЬТУР

Группе по управлению энергией на заводе, выпускающем продукты из зерновых культур, было поручено найти пути для снижения энергопотребления. Причина заключалась в том, что технологические процессы, используемые для переработки зерна, периодически потребляли значительные объемы перегретой горячей воды[1].

Этот проект требовал сотрудничества между инженером-технологом Мэттом и руководителем проекта по управлению энергопотреблением Лораном. Инженерам-технологам и команде по управлению энергопотреблением была поставлена задача — найти меры по энергосбережению.

Котлы для приготовления продукта потребляли значительные объемы горячей воды для поддержания надлежащих температур. Для оптимизации энергопотребления был предложен новый коллектор горячей воды с программным обеспечением для управления подачей пресной воды. При его использовании, благодаря более жестким стратегиям контроля температуры, количество жидкости, сливаемой из предыдущей партии, можно было уменьшить, чтобы минимизировать количество горячей воды при одновременном поддержании требуемой температуры варки зернового сырья. В процессе изготовления конечного продукта экономия энергии будет достигаться за счет уменьшения объема новых добавок перегретой воды в автоклав.

Для того чтобы разработать управляемое данными решение вышеуказанной проблемы, Мэтт экспортировал данные пакетной обработки из исторических данных процесса и системы его выполнения в отдельные электронные таблицы. Была установлена новая система управления, поэтому Мэтт отказался от данных старой системы и начал новый анализ. Объем информации ограничивал возможности ее экспорта и требовал сужения анализа с года до последнего квартала. Кроме того, проблемой было несоответствие данных временных рядов историческим данным и текущей системе производства.

Из-за этих различий требовалась ручная синхронизация времени между двумя системами. Используя «чапаевский метод», Мэтт из нескольких пакетов создал сжатое представление данных с низким разрешением и уже тогда смог экстраполировать результаты в течение года.

Несмотря на все усилия Мэтта по сбору имеющихся данных о периодическом приготовлении конечной продукции, результаты были неоднозначными. Новые таблицы все еще оставались слишком большими, чтобы их можно было легко разделить с командой по управлению энергией. Из-за размера и сложности таблицы периодически падали после включения расчетов и диаграмм. Мэтт и Лорен потратили значительное время на просеивание и сортировку данных, чтобы получить то, что могло бы дать практически

РИС. 4. ▲ Замена анализа на основе электронных таблиц передовым аналитическим программным обеспечением позволила команде по управлению энергопотреблением быстро получать и обмениваться информацией

[1] Перегретая вода (англ. super-heated hot water) — это жидкая вода под давлением при температуре между обычной температурой кипения =100 °C и критической температурой +374 °C, также известна как «докритическая вода», или «горячая вода под давлением». — Прим. пер.

полезные результаты, как показано на рис. 3.

Быстрый обзор сделанного показал, что Мэтт потратил большую часть своего времени на очистку, сортировку, просеивание, копирование и вставку представляющих интерес данных в электронные таблицы. Причем он израсходовал относительно мало времени на фактический анализ предлагаемых оперативных изменений, хотя и эта проблема все еще оставалась довольно обременительной. В ходе выполнения проекта Мэтт был повышен до новой должности, и решение этой задачи взял на себя новый инженер. В итоге различные методы подготовки данных, использованные двумя технологами, привели к расхождению результатов.

Из описания уже этого единичного случая мы ясно видим проблемы, с которыми сталкиваются предприятия при анализе данных технологических и производственных процессов, представленных с помощью электронных таблиц. Несмотря на все свои усилия, группа технологов и команда по управлению энергопотреблением постоянно повторяли одни и те же шаги в обработке данных. Проблема заключалась не в нехватке данных — как правило, ее нет ни на одном промышленном предприятии. Скорее это были данные, не синхронизированные по времени и найденные в разных местах и представленные в разных форматах. Использование электронных таблиц для обмена данными с различными группами также оказалось трудным делом, а эффективная визуализация и создание отчетов были практически невозможны.

РЕШЕНИЕ ПРОБЛЕМ — ПРИМЕНЕНИЕ РАСШИРЕННОЙ АНАЛИТИКИ ДАННЫХ

Расширенное современное программное обеспечение для анализа данных имеет доступ к данным там, где они находятся. Копирование и вставка здесь уже не требуются, потому что программное обеспечение объединяет данные высокого разрешения из нескольких источников (рис. 4). Простые команды запросов облегчают нацеливание данных и поддерживают их упорядоченную сортировку, очистку и сборку, причем только тех из них, что поступали от серверов хранения исторических данных и других источников.

Программное обеспечение для расширенной аналитики также поддерживает базовые вычисления и другие математические функции, которые инженеры используют для преобразования данных в визуальное суммирование трендов и других соответствующих видов данных. Интересующие сведения легко собираются и контекстуализируются для моделирования поведения конкретного технического или производственного процесса в будущем — например, для прогнозного технического обслуживания. При этом, что немаловажно, объем обучения новых пользователей здесь минимален.

Программные средства расширенной аналитики ускоряют процессы очистки, визуализации, поиска, контекстуализации и моделирования данных. Используя такие инструменты, инженеры сосредотачиваются именно на сборе знаний, совместной работе и решении проблем, а не на разгадывании полученных данных. ⬢