

Open Source Software,
Люди, и Сообщества
как стратегия выживания

Артеми́й Козы́рь



Что я делаю:

- [Analytics Engineering @ Wheely](#)
- [Performance Marketing Analytics](#)
- Blogging: [Technology Enthusiast](#) and [Habr](#)
- ex DE & DWH @ Sber, Sibur, Moscow Exchange

С чем работаю:

- DWH: [Redshift](#), Snowflake, Clickhouse
- Transformations: [Data Build Tool](#)
- BI: [Looker](#), Metabase, Datalens
- [SQL](#), Python, Bash, Docker, Github Actions
- AWS, Yandex.Cloud

Все суждения в этой презентации -
моё личное мнение и домыслы

Ключевая идея #1:

Поздно делать своё ПО.

Рынок ПО РФ проиграл и отстал.

Годы (десятилетия) инвестиций в западное ПО привели к стратегическому поражению

- Непрозрачные решения в пользу малой группы заинтересованных лиц с сомнительными результатами
- Выбор западных вендоров - это всегда инвестиции в развитие "их" рынка
- Ликвидация и лишение конкурентоспособности собственного рынка ПО (в долгосрочной перспективе)

Рынок ПО РФ зависим и бесперспективен

Можно выделить несколько групп Отечественного ПО:

- Упаковка Open Source решений (мало) – *Arenadata*
- Нишевые приемлемые решения, с которыми можно работать (мало) – *Yandex Datalens, Yandex Metrika*
- Бесперспективный, отставший и почти никому не нужный софт (много)

Задача: придумайте архитектуру аналитического решения с использованием отечественного ПО

Конкуренция с западным ПО на равных: возможности, поддержка, стоимость (лицензии), гибкость

- **Extract-Load:** синк данных источников на регулярной основе без потерь и дублирования
- **Transform:** трансформация исходных данных в осмысленные витрины
- **BI:** визуализация и интерактивная аналитика, дашбординг
- **Advanced analytics:** A/B, ML, reverse ETL, real-time scoring / fraud monitoring

Ключевая идея #2:

Выжить можно, если использовать
Open Source Software

Основные тренды в управлении данными сегодня: Extract-Load

- Надежная и стабильная репликация данных на регулярной основе без потерь и дублирования
- Простейшие трансформации (пропуск, обезличивание, хеш-сумма атрибута)

Open Source Solutions:

- <https://github.com/airbytehq/airbyte>
- <https://github.com/dagster-io/dagster>
- <https://github.com/jitsucom/jitsu>

Основные тренды в управлении данными сегодня: Databases

- Performance, Scalability, Fault-tolerance
- Columnar formats, Data compression, Partitioning, Clustering
- Data integration: External Data, Federated Queries
- Advanced functions support: JSON, Geospatial, ML

Open Source Solutions:

- <https://github.com/ClickHouse/ClickHouse>
- <https://github.com/trinodb/trino>
- <https://github.com/duckdb/duckdb>

Основные тренды в управлении данными сегодня: Transform

- Моделирование, Развертывание, Тестирование, Документирование Витрин Данных
- Версионирование, шаблонизация, переиспользование (модулярность) кода
- Независимость от используемой СУБД (Database agnostic)

Open Source Solutions:

- <https://github.com/dbt-labs/dbt-core>

Основные тренды в управлении данными сегодня: Business Intelligence

- Декларативный и версионизируемый подход к визуализациям и дашбордам
- Взаимодействие через программный интерфейс (API)
- Headless BI, Semantic Layer

Open Source Solutions:

- <https://github.com/apache/superset>
- <https://github.com/metabase/metabase>
- <https://github.com/cube-js/cube>

Основные тренды в управлении данными сегодня: Advanced analytics

- A/B + Statistics
- ML + Deep Learning + MLOps
- Reverse ETL (DWH to Operational Systems)
- Real-time scoring / fraud monitoring / bidding

Крупнейшие ТНК – контрибуторы популярных Open Source проектов ([ASF](#))



Какой вклад в рынок отечественного ПО внесли топовые российские компании?

#	Название	Чистая прибыль в 2021 году	Чистая прибыль в 2020 году
1	 Газпром Нефть и газ	2159.1 млрд рублей	162.4 млрд рублей
2	 Сбербанк Финансы	1245.9 млрд рублей	760.3 млрд рублей
3	 Роснефть Нефть и газ	1012 млрд рублей	166 млрд рублей
4	 Лукойл Нефть и газ	775.5 млрд рублей	16.6 млрд рублей
5	 Сургутнефтегаз Нефть и газ	513.2 млрд рублей	729.6 млрд рублей
6	 Норильский никель Цветная металлургия	512.9 млрд рублей	263.8 млрд рублей
7	 Новатэк Нефть и газ	451.6 млрд рублей	78.6 млрд рублей
8	 НЛМК Черная металлургия	371.7 млрд рублей	91 млрд рублей

Какой вклад в рынок отечественного ПО внесли топовые российские технологические компании?

№ ↕	Название ↕	Место по итогам 2020 года	Местоположение центрального офиса	Суммарный объем выручки от ИКТ-деятельности (без учета НДС, тыс. руб.), 2021 год*	Суммарный объем выручки от ИКТ-деятельности (без учета НДС, тыс. руб.), 2020 год*	Темп роста (%) ↕	Численность специалистов (чел.), 2021 год (** показатель численности всех сотрудников)	Изменение численности спец. за год
1	Softline	2	Москва	152189784	123092003	23.6	6 000**	1100
2	Т1	-	Москва	76561499	50570828	51.4	8 789**	4328
3	МТС	-	Москва	51639215	43971563	17.4	5555	407
4	АйТеко	4	Москва	35376438	29098064	21.6	3 000**	-50
5	«КРОК»	3	Москва	34869550	32979335	5.7	2 192**	208
6	IBS	-	Москва	28306448	30028265	-5.7	2758	-115
7	ЦФТ («Центр финансовых технологий»)	6	Москва	26400331	24277306	8.7	3 934**	657
8	Sitronics Group	-	Москва	24298373	20741627	17.1	1 529**	531
9	ФОРС	7	Москва	16416156	17856271	-8.1	379	-66
10	ICL-КПО ВС	9	Казань	15689981	13487090	16.3	3952	410

Ключевая идея #3:

Лучшая инвестиция сегодня - в умы и руки (которые утекают из РФ)

Без людей ПО ничего не значит

- ПО не существует без тех, кто способен создавать, поддерживать и развивать его
- Надежда на специалистов и команды внедрения (ум и головы, которые пока есть в РФ)

Чтобы удерживать умы и специалистов, необходимо создавать комфортные условия

- Стабильность и долгосрочное планирование
- Инвестиционный и политический климат
- Условия труда
- Развитие сообществ специалистов, площадок для обмена идеями

Будут условия – будет и развитие

- *Сегодня* внедряем Open Source Software
- *Сегодня* учимся, оттачиваем навыки, развиваемся
- *Завтра* модифицируем ПО, делаем forks, доработки, что-то своё
- *Послезавтра* достигаем того, уровня, когда наши решения будут выигрывать конкуренцию "у них"
- *Послезавтра* специалисты со всего мира хотят жить и работать в России

Люди организуются в сообщества для общения и поиска решений

The screenshot displays a Slack workspace named "MergeTree" with 7,905 members. The left sidebar shows a list of recent messages from various users, including Slach, Denny, Alexander Tokmakov, and others. The main chat area shows a conversation between several users, including Valentin Shultsev and Alexander Tokmakov, discussing the storage of MergeTree data. The chat background is a green patterned wallpaper. The interface includes search bars, a pinned message, and a "Translate to English" button.

Search messages in

ClickHouse не тормозит

Found 4048 messages

Slach [altinity] 15:21
он пытается part перенести 4. DB::MergeTreeData::moveParts a...

Denny [Altinity] 03:56
<https://kb.altinity.com/engines/mergetree-table-engine-family/pick-...>

Alexander Tokmakov 03:37
Партиция устроена сильно сложнее, чем один файл. Но даже ес...

Denny [Altinity] Mon
ну в КХ все неочевидно. Например insert_deduplication можно ...

Ivan Leventsov Sun
#materialized_view #lost #потеря_данных привет! у нас есть та...

Slach [altinity] Sat
а какие кстати? Они на уровне диска или на уровне BACKUP / R...

Denny [Altinity] Fri
Короче КХ использует не индексы, а тот факт что сортировка сор...

Sergey Cherkashin Thu
Господа, а подскажите, пожалуйста, как можно удалить матери...

Andrey Prohorov Thu
Всем приветы, подскажите пожалуйста, возможно в конфиге us...

Denny [Altinity] Wed
вы ниче не поняли. у вас был инсерт в Distributed таблицу, он ...

Sergey Redyuk Wed
День добрый! Коллега (не русскоговорящий) столкнулся со след...

Slach [altinity] Wed
<https://clickhouse.com/docs/en/engines/table-engines/mergetree-f...>

Slach [altinity] Wed

ClickHouse не тормозит
7 905 members

Pinned message
Clickhouse — база данных для гибкого анализа огромных потоков информации. Microsoft, Uber и Deutsche Bank — лишь некоторые известные пользователи. Как эксперим...

Translate to English

SOURCE (... USER 'default'), которого не существовало.
edited 22:55

June 13

Valentin Shultsev
Всем привет. А сколько рекомендовано хранить строк в 1 партиции?
1 03:30

Alexander Tokmakov Reply
Valentin Shultsev
Всем привет. А сколько рек...
03:34

Valentin Shultsev
То есть миллион строк лучше хранить 1 файле, а не разбивать на более мелкие?
edited 03:35

Alexander Tokmakov admin
Партиция устроена сильно сложнее, чем один файл. Но даже если разбивать, то в итоге они всё равно помёрзнутся в один большой парт, так работает MergeTree
03:37

Ну то есть между партициями ничего не мёрзится, но смысла разбивать данные на партиции всё равно особо нет, кроме случаев когда нужно делать какие-то операции с отдельными партициями (например, обычно делают партиционирование по месяцам, чтобы можно было легко удалить старые партиции)
03:40

Valentin Shultsev
В целом понял, спасибо.
1 03:42

Denny [Altinity]
Valentin Shultsev
В целом понял, спасибо.
<https://kb.altinity.com/engines/mergetree-table-engine-family/pick-keys/#partition-by>
Altinity

Write a message...

Ключевые выводы

Рынок ПО РФ локальный, тусклый и отсталый

- Импортозамещения не было, нет, и не будет
- Чтобы двигаться вперед, нужно сначала проводить работу над ошибками

Чтобы выживать, нужно становиться частью бОльшего сообщества

- Используйте OSS, добавляйте новые фичи, исправляйте баги

Чтобы людям, создающим ПО было комфортнее жить и делать это в РФ, необходимо создавать среду

- И тогда, возможно, через десяток лет в РФ начнет появляться софт, востребованный в мировом масштабе

Спасибо. Вопросы?



@enthusiastech



Artemiy Kozyr



kzzzr