From Engineer to Architect – Engineering Disciplines for High-Quality and Secure Code

鉴释 xcalibyte

Shinming Liu

Chief Architect, Xcalibyte

Who Am I

刘新铭 Shin-Ming Liu

Chief Architect & Co-founder, Xcalibyte

- Compiler Scientist
- Former Director, Intel IOT Research Lab, China
- Former Director, HP Compiler Technology Lab
- 10+ patents granted in program analysis and compiler optimization

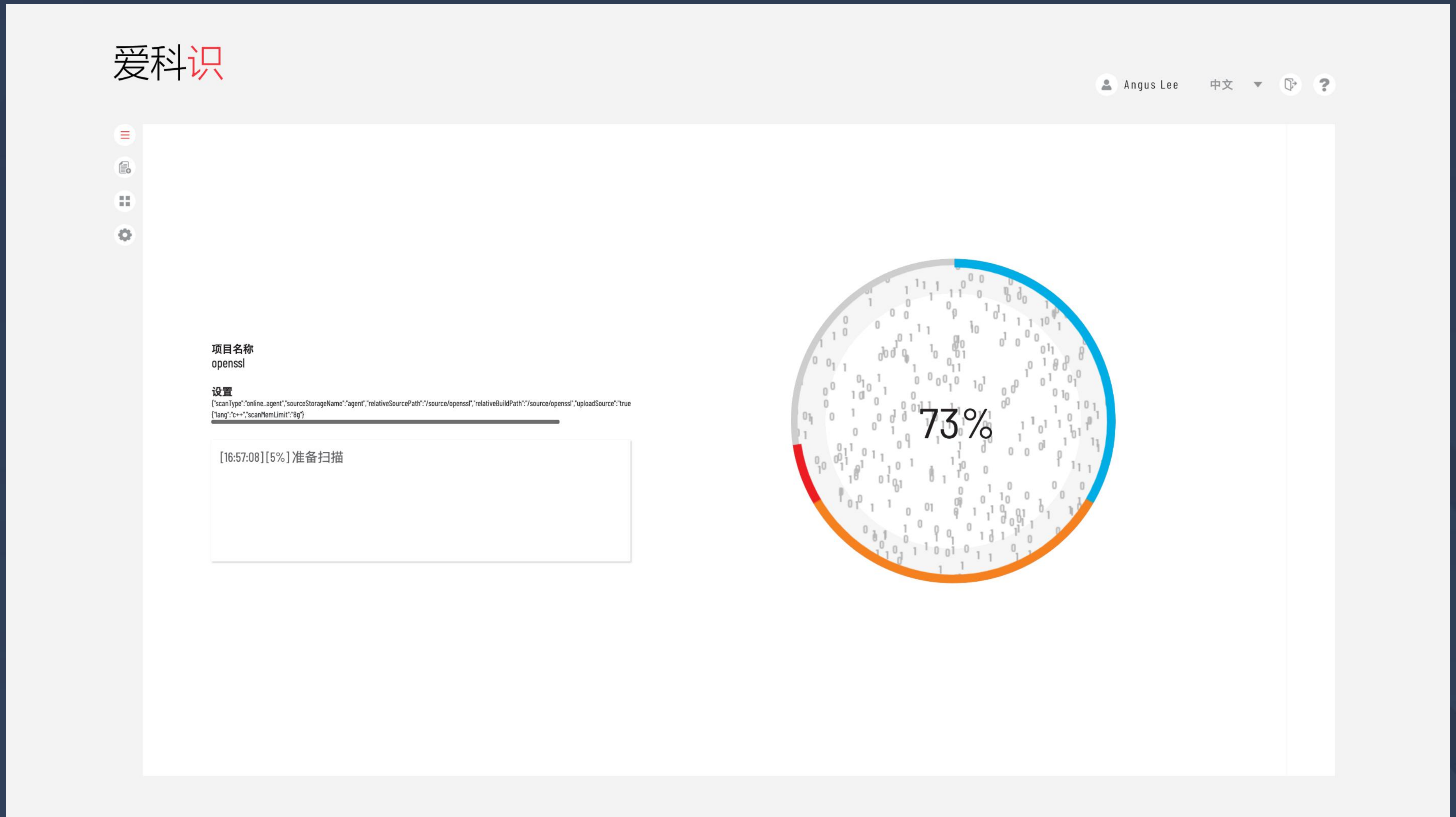
Agenda

- 案例：数据库重构项目
- 用 First Principle 来分析，思考
- 用性能数据来佐证、决策，并管控
- 用程序内置工具来监控架构劣坏
- 为算法适配程序语言
- 重构要及时并持续进行
- 自我提升

CASE — DATABASE REFACTOR PROJECT

Pain Points

- Service Does Not Complete
- Service Timeout
- Random Service Restart
- UI Response Time Takes More than 25min for Extreme Cases



项目名称
openssl

设置

```
{"scanType":"online_agent","sourceStorageName":"agent","relativeSourcePath":"/source/openssl","relativeBuildPath":"/source/openssl","uploadSource":"true
{"lang":"c++","scanMemLimit":"8g"}
```

```
[16:57:08][5%]准备扫描
```

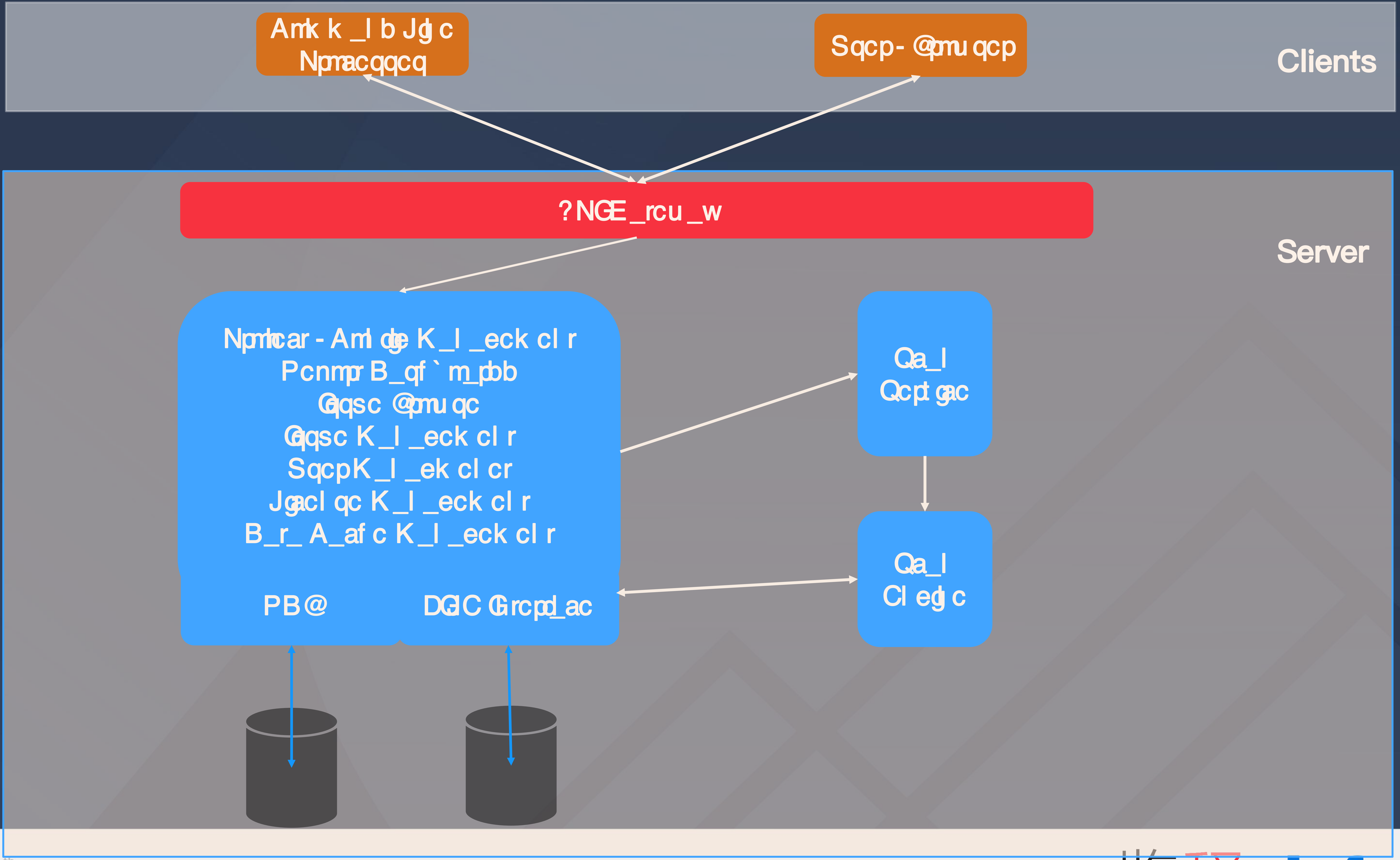
Xcalscan was a Monolithic Java Application

AGAB8
EgJ_`*
Hcl i gq*
Rp_t gq AG
,,,

GBCq8
Cajgnqc*
TQ Ambc*

QAK 8
Eg*

Amk k Rmmjq8
Ck _gj*



Amk k _l b Jgj c
Npmacqqcq

Sqcp- @pmu qcp

Clients

?NGE_rcu_w

Server

Npmhcar - Aml dge K_l _eck cl r
Pcnmpr B_qf ` m_pbb
Gqsc @pmu qc
Gqsc K_l _eck cl r
SqcpK_l _ek cl cr
Jgacl qc K_l _eck cl r
B_r_ A_afc K_l _eck cl r

PB@

DGIC Grcpd_ac

Qa_l
Qcpt gac

Qa_l
Cl egj c

PERFORMANCE DATA COLLECTED

Measurement Made:

- ➢ Input Data Set Size
- ➢ DB Tables Size Collected For The Input
- ➢ Memory Footprint and Elapse Time
  - ❑ Data Ingestion into DB for the Input
  - ❑ IssueChangeAnalysis function for One Data Set
- ➢ Response Time for Dashboard Display

PERFORMANCE DATA COLLECTED

Measurement Made for MySQL5.7

- ➢ Input Data Set Size – 2.2 million LOC, 565MB Issue File
- ➢ DB Tables Size Collected For The Input – 2GB
- ➢ Memory Footprint and Elapse Time
    - ❑ Data Ingestion into DB for The Input – 22GB
    - ❑ IssueChangeAnalysis function for One Data Set – 12G
- ➢ Response Time for Dashboard Display – 180 Sec

ISSUE IDENTIFIED

- Direct Correlation
  - ➢ Input Data Set Size ➔ Memory Footprint ➔ Elapse Time
- DB Tables Size Increases Accumulatively
  - ➢ This result in Slow Response Time DB Query
- Inefficient Query Function Used
  - ➢ API Involves Full Table Scan, Join, Select
- Interpretive Language Used in IssueChangeAnalysis Algorithm

DESIGN CHANGES MADE

- Redesign Data Format for Data Produced by Scan Engines
  - Use String Table – Eliminate Duplicate String
  - Use Proprietary Data Format – Internal Use Only
- Database Table Changes:
  - Source Code Issue Introduction Time (GIT commit ID)
  - Ingest New Issue Introduced Only, Minimize DB Size Increase
  - Remove Read-only Data from RDB
- Precompute Time Consuming Frequent Query
- Use C++ for Complex Algorithms

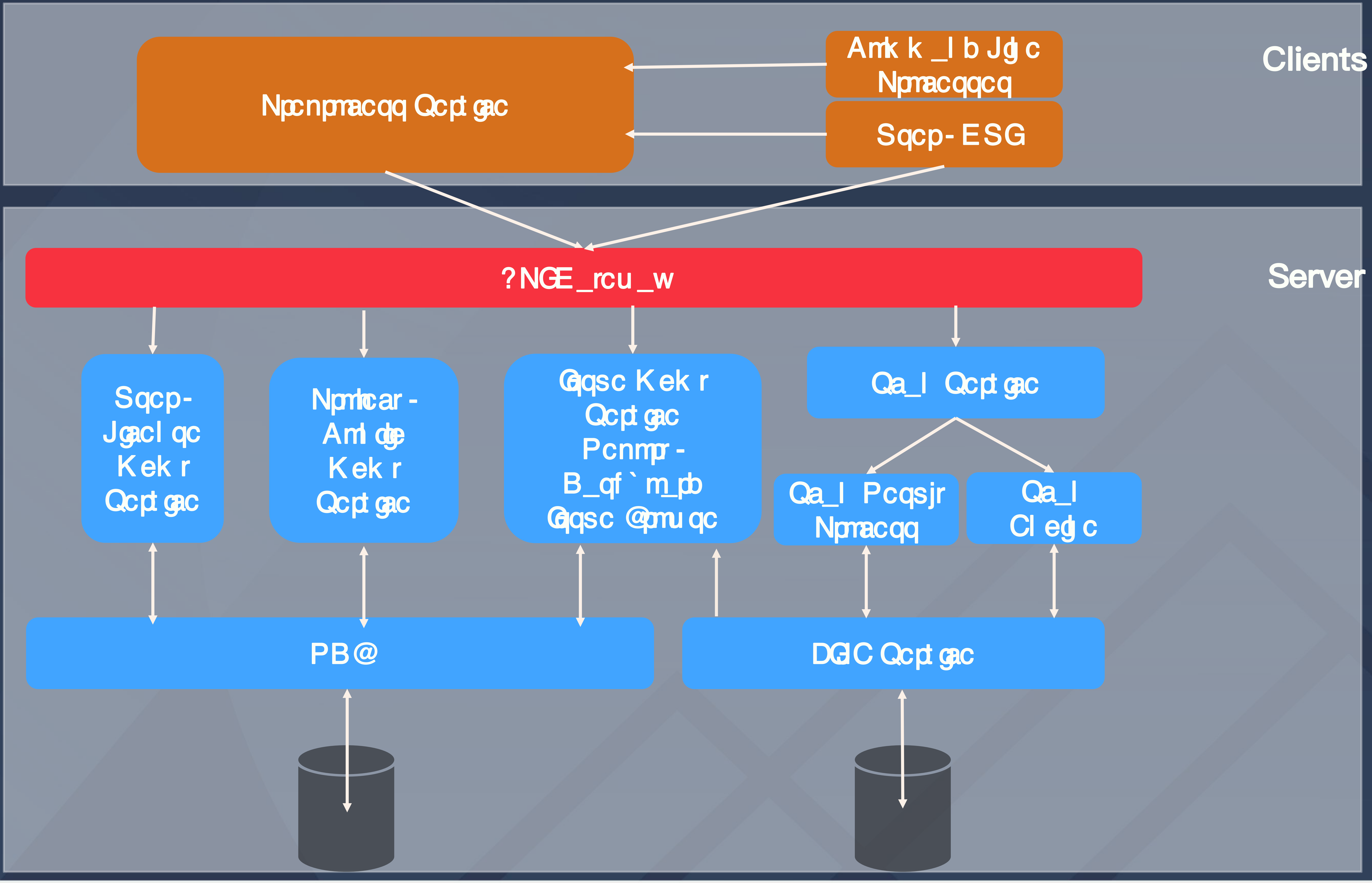
Xcalscan Architecture After Refactor

AGAB8
EgJ_`*
Hcl i gj q*
Rp_t gj AG
,,,

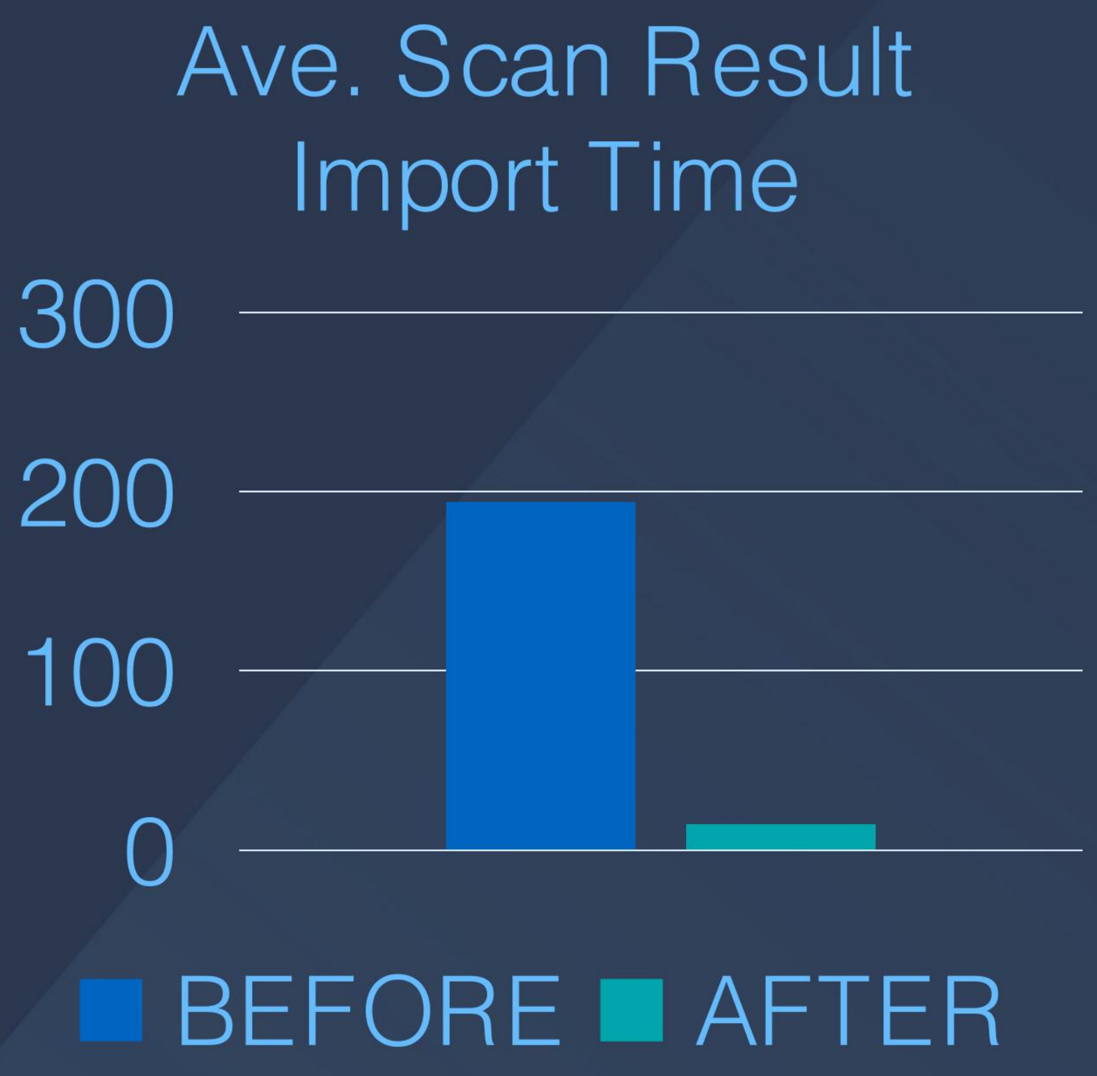
GBCq8
Cajgnqc*
TQ Ambc*

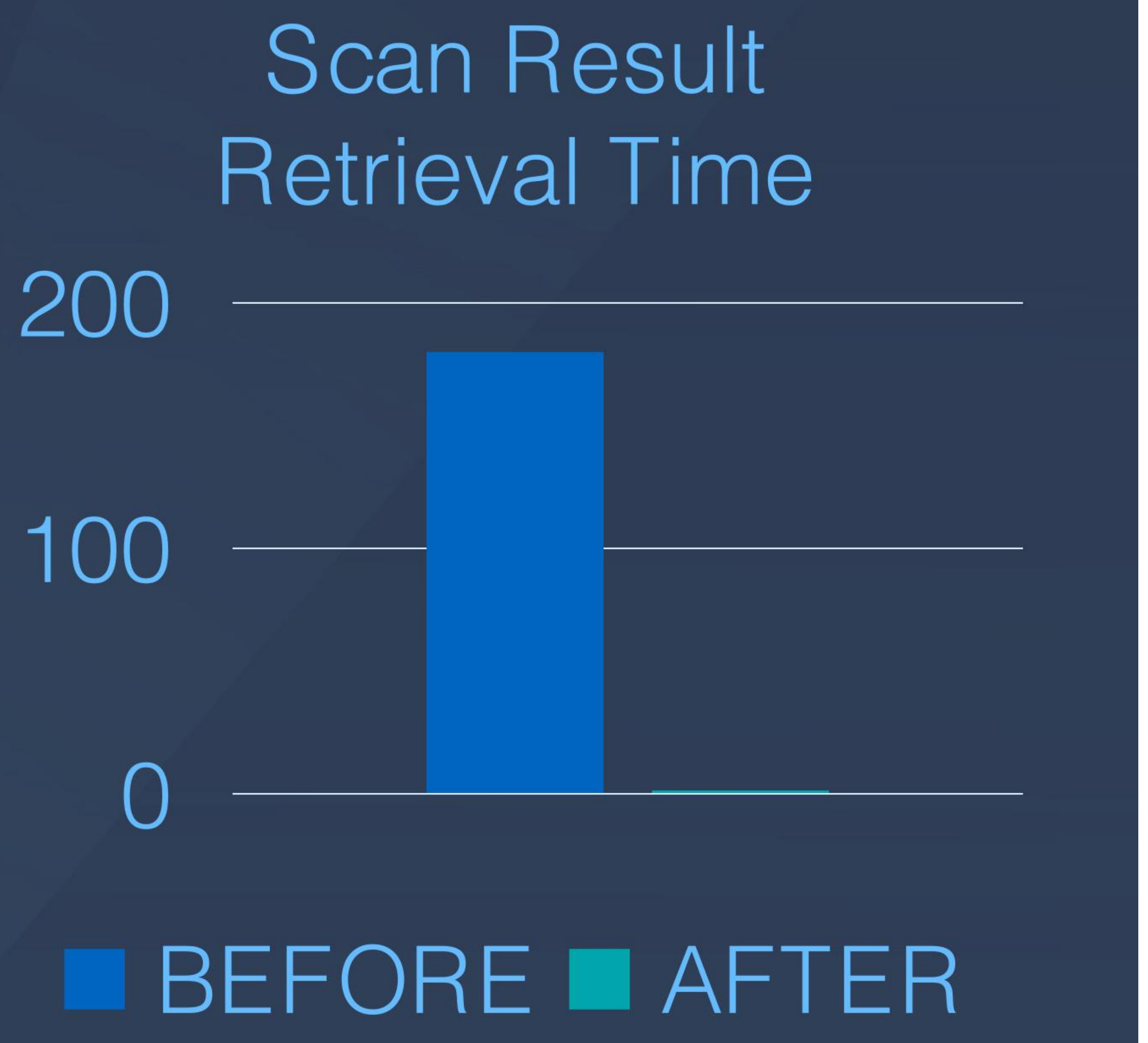
QAK8
Eg*

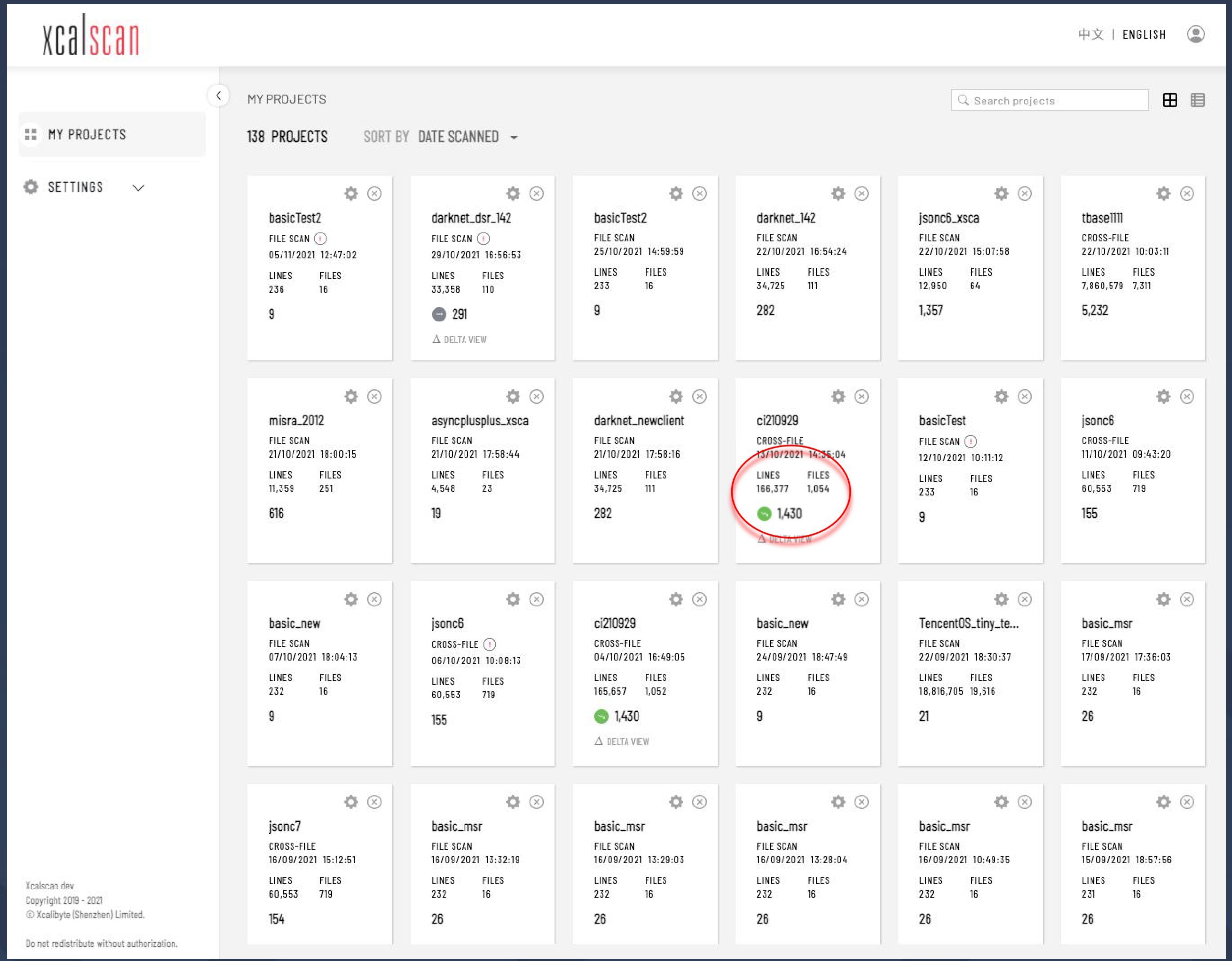
Amk k Rmmjq8
Ck _gj*



PERFORMANCE IMPROVEMENTS







138 PROJECTS SORT BY DATE SCANNED

basicTest2
FILE SCAN
05/11/2021 12:47:02
LINES 236 FILES 16
9

darknet_dsr_142
FILE SCAN
29/10/2021 16:56:53
LINES 33,358 FILES 110
291
DELTA VIEW

basicTest2
FILE SCAN
25/10/2021 14:59:59
LINES 233 FILES 16
9

darknet_142
FILE SCAN
22/10/2021 16:54:24
LINES 34,725 FILES 111
282

jsonc6_xsca
FILE SCAN
22/10/2021 15:07:58
LINES 12,950 FILES 64
1,357

tbase1111
CROSS-FILE
22/10/2021 10:03:11
LINES 7,860,579 FILES 7,311
5,232

misra_2012
FILE SCAN
21/10/2021 18:00:15
LINES 11,359 FILES 251
616

asyncplusplus_xsca
FILE SCAN
21/10/2021 17:58:44
LINES 4,548 FILES 23
19

darknet_newclient
FILE SCAN
21/10/2021 17:58:16
LINES 34,725 FILES 111
282

ci210929
CROSS-FILE
13/10/2021 14:35:04
LINES 166,377 FILES 1,054
1,430
DELTA VIEW

basicTest
FILE SCAN
12/10/2021 10:11:12
LINES 233 FILES 16
9

jsonc6
CROSS-FILE
11/10/2021 09:43:20
LINES 60,553 FILES 719
155

basic_new
FILE SCAN
07/10/2021 18:04:13
LINES 232 FILES 16
9

jsonc6
CROSS-FILE
06/10/2021 10:08:13
LINES 60,553 FILES 719
155

ci210929
CROSS-FILE
04/10/2021 16:49:05
LINES 165,657 FILES 1,052
1,430
DELTA VIEW

basic_new
FILE SCAN
24/09/2021 18:47:49
LINES 232 FILES 16
9

TencentOS_tiny_te...
FILE SCAN
22/09/2021 18:30:37
LINES 18,816,705 FILES 19,616
21

basic_msr
FILE SCAN
17/09/2021 17:36:03
LINES 232 FILES 16
26

jsonc7
CROSS-FILE
16/09/2021 15:12:51
LINES 60,553 FILES 719
154

basic_msr
FILE SCAN
16/09/2021 13:32:19
LINES 232 FILES 16
26

basic_msr
FILE SCAN
16/09/2021 13:29:03
LINES 232 FILES 16
26

basic_msr
FILE SCAN
16/09/2021 13:28:04
LINES 232 FILES 16
26

basic_msr
FILE SCAN
16/09/2021 10:49:35
LINES 232 FILES 16
26

basic_msr
FILE SCAN
15/09/2021 18:57:56
LINES 231 FILES 16
26

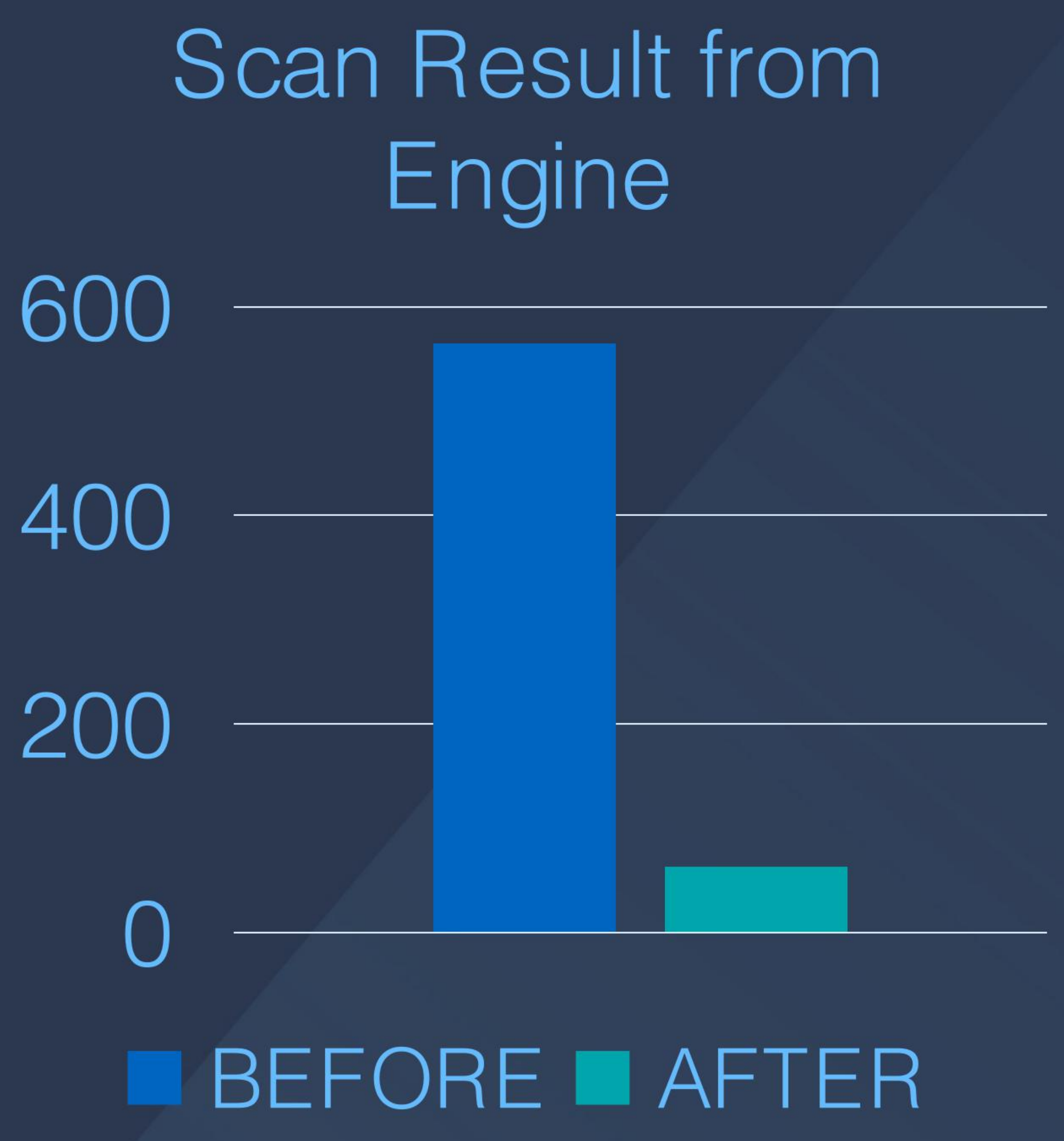
Xcalscan dev
Copyright 2019 - 2021
© Xcalibyte (Shenzhen) Limited.

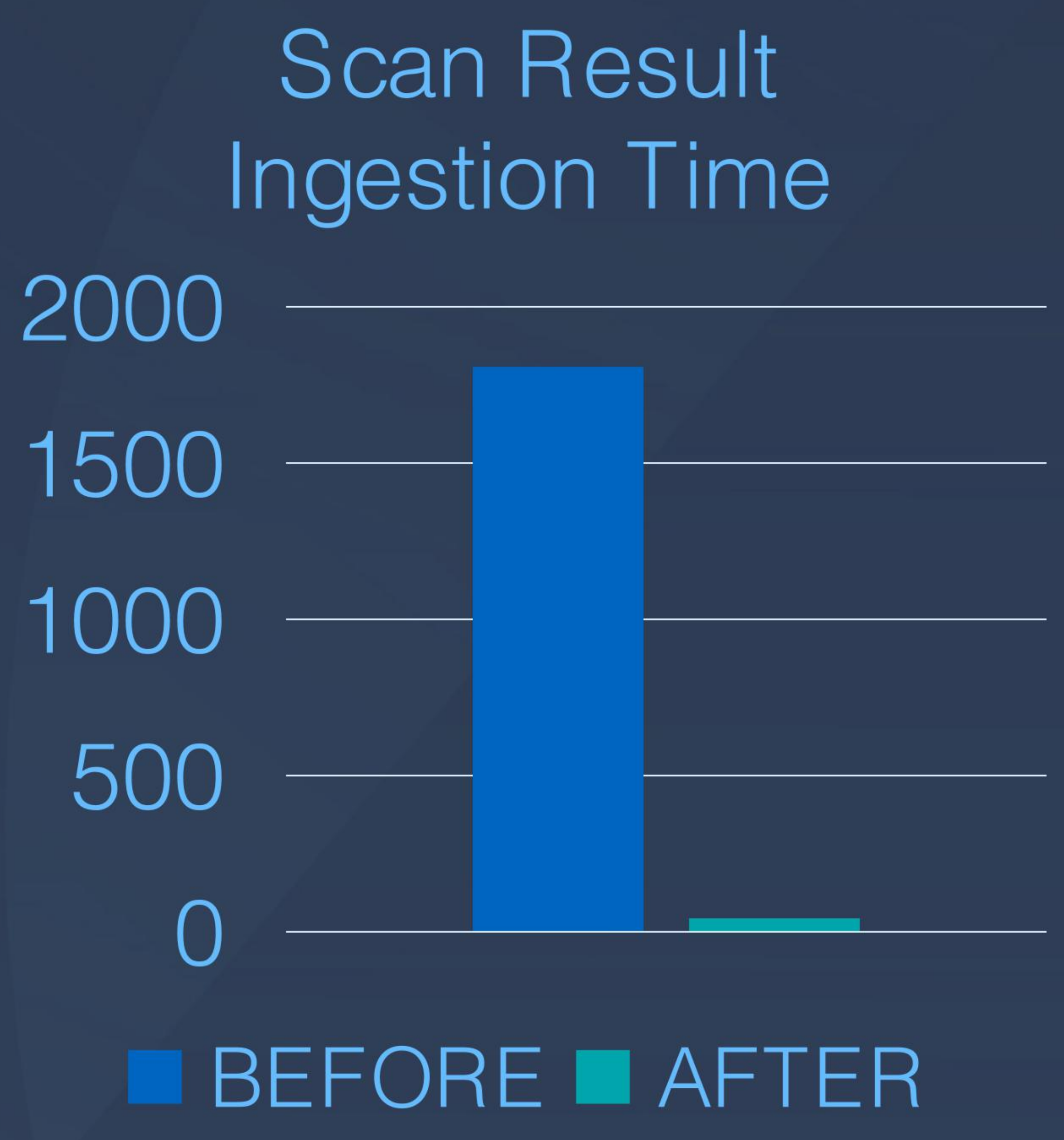
Do not redistribute without authorization.

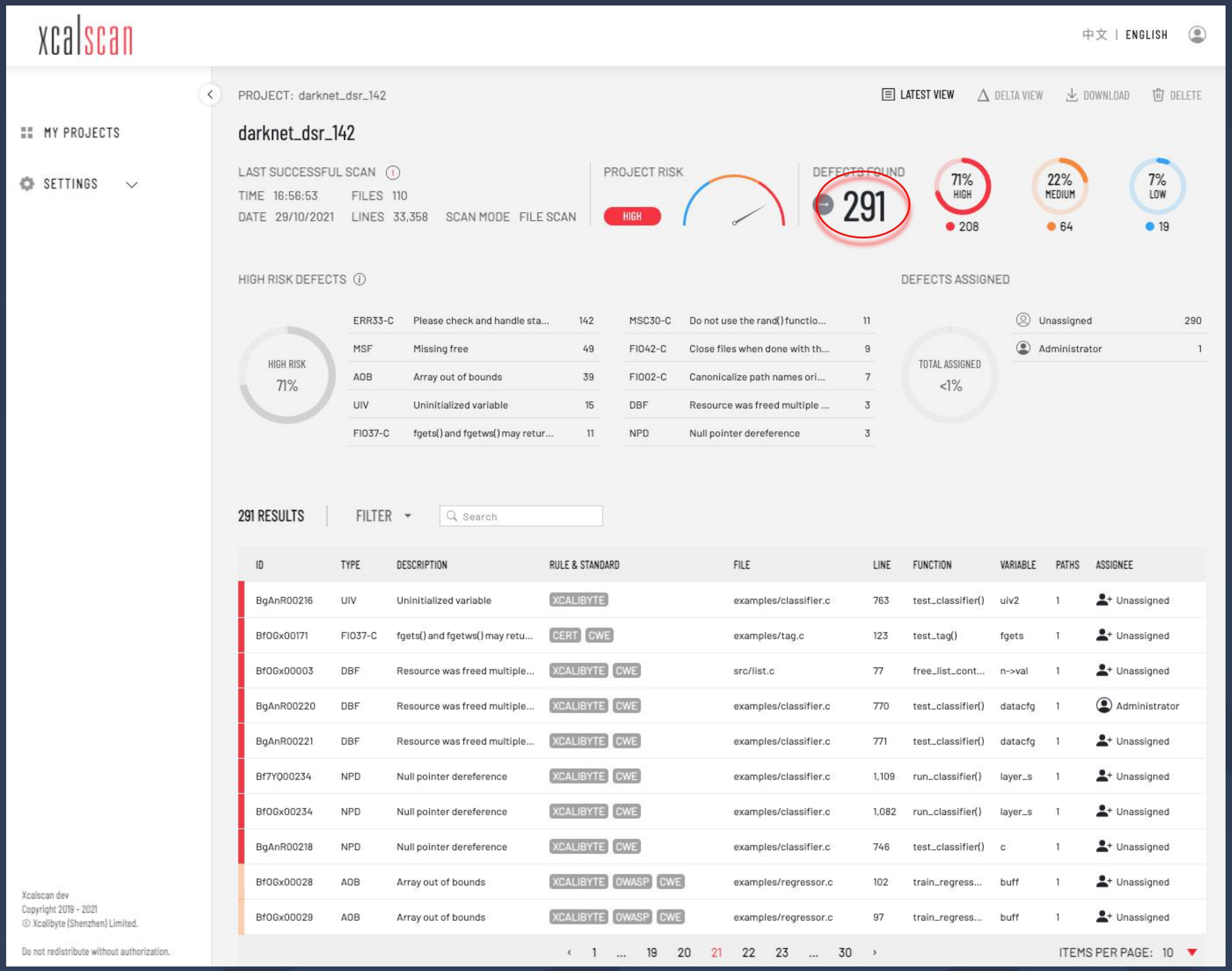
- Precompute Summary Data, Reduce UI Dashboard Load Time

PERFORMANCE IMPROVEMENTS

(cont)


Scan Result from Engine


Scan Result Ingestion Time



PROJECT: darknet_dsr_142

darknet_dsr_142

LAST SUCCESSFUL SCAN
TIME 16:56:53 FILES 110
DATE 29/10/2021 LINES 33,358 SCAN MODE FILE SCAN

PROJECT RISK HIGH

DEFECTS FOUND 291

71% HIGH 208 · 22% MEDIUM 64 · 7% LOW 19

HIGH RISK DEFECTS

HIGH RISK 71%

ERR33-C	Please check and handle sta...	142	MSC30-C	Do not use the rand() functio...	11
MSF	Missing free	49	FIO42-C	Close files when done with th...	9
AOB	Array out of bounds	39	FIO02-C	Canonicalize path names ori...	7
UIV	Uninitialized variable	15	DBF	Resource was freed multiple ...	3
FIO37-C	fgets() and fgetws() may retur...	11	NPD	Null pointer dereference	3

DEFECTS ASSIGNED

TOTAL ASSIGNED <1%

Unassigned	290
Administrator	1

291 RESULTS | FILTER | Search

ID	TYPE	DESCRIPTION	RULE & STANDARD	FILE	LINE	FUNCTION	VARIABLE	PATHS	ASSIGNEE
BgAnR00216	UIV	Uninitialized variable	XCALIBYTE	examples/classifier.c	763	test_classifier()	uiv2	1	Unassigned
BfOGx00171	FIO37-C	fgets() and fgetws() may retu...	CERT CWE	examples/tag.c	123	test_tag()	fgets	1	Unassigned
BfOGx00003	DBF	Resource was freed multiple...	XCALIBYTE CWE	src/list.c	77	free_list_cont...	n->val	1	Unassigned
BgAnR00220	DBF	Resource was freed multiple...	XCALIBYTE CWE	examples/classifier.c	770	test_classifier()	datacfg	1	Administrator
BgAnR00221	DBF	Resource was freed multiple...	XCALIBYTE CWE	examples/classifier.c	771	test_classifier()	datacfg	1	Unassigned
Bf7YQ00234	NPD	Null pointer dereference	XCALIBYTE CWE	examples/classifier.c	1,109	run_classifier()	layer_s	1	Unassigned
BfOGx00234	NPD	Null pointer dereference	XCALIBYTE CWE	examples/classifier.c	1,082	run_classifier()	layer_s	1	Unassigned
BgAnR00218	NPD	Null pointer dereference	XCALIBYTE CWE	examples/classifier.c	746	test_classifier()	c	1	Unassigned
BfOGx00028	AOB	Array out of bounds	XCALIBYTE OWASP CWE	examples/regressor.c	102	train_regress...	buff	1	Unassigned
BfOGx00029	AOB	Array out of bounds	XCALIBYTE OWASP CWE	examples/regressor.c	97	train_regress...	buff	1	Unassigned

‹ 1 ... 19 20 21 22 23 ... 30 › ITEMS PER PAGE: 10

Xcalscan dev
Copyright 2019 - 2021
© Xcalibyte (Shenzhen) Limited.

Do not redistribute without authorization.

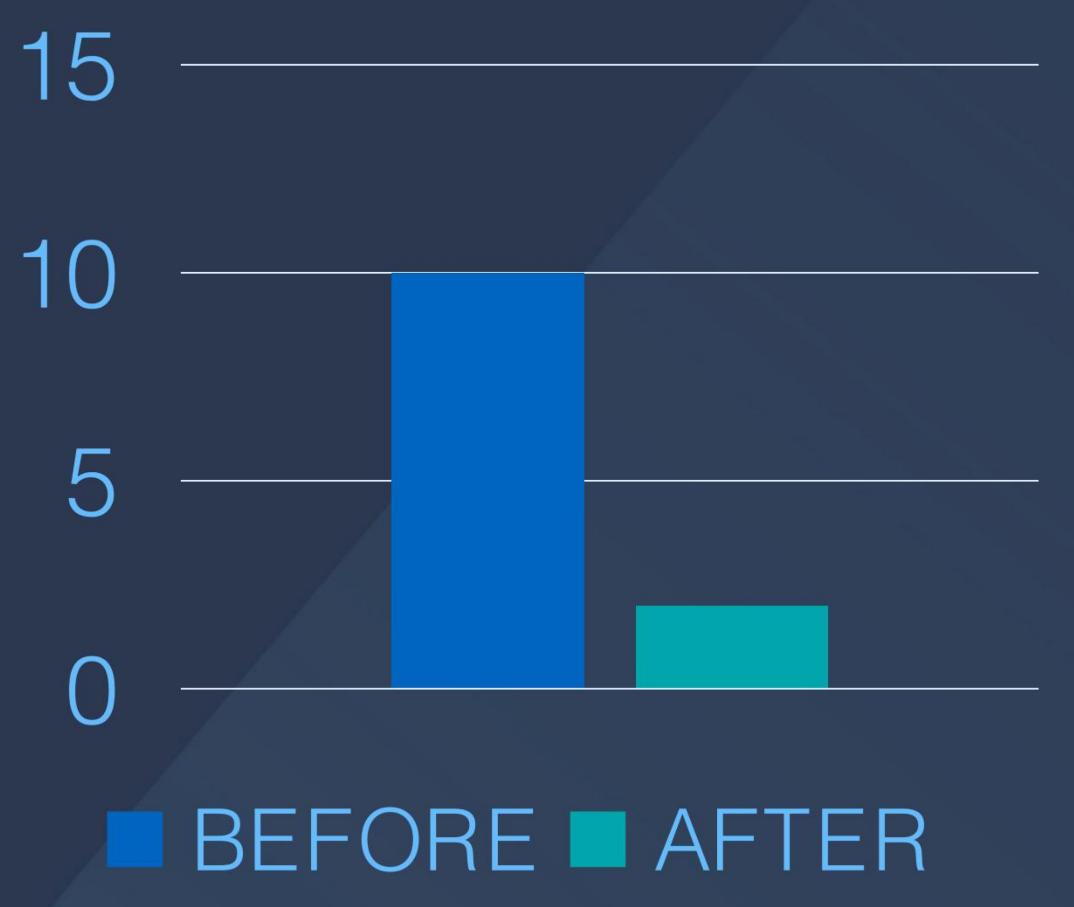
- Identify and Skip Redundant Data, Enables Large Project Scan Such as MySQL5.7

PERFORMANCE IMPROVEMENTS

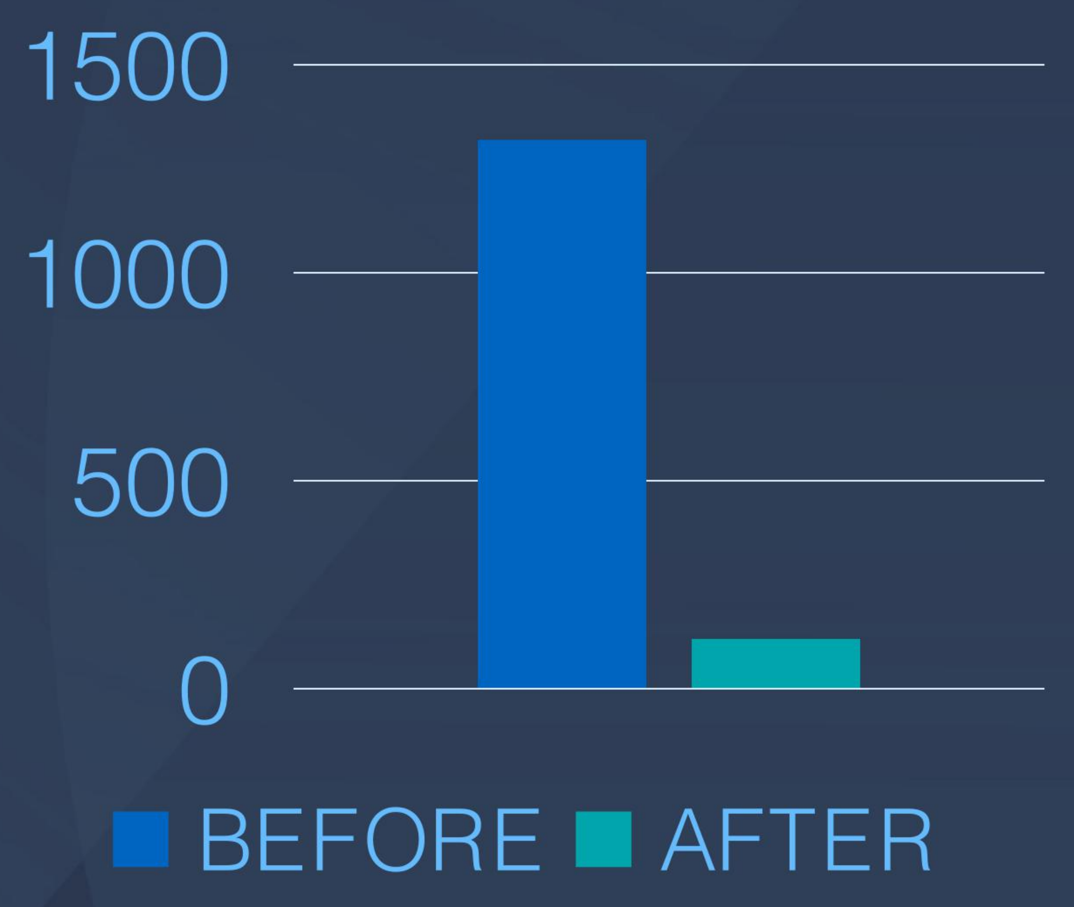
(cont)

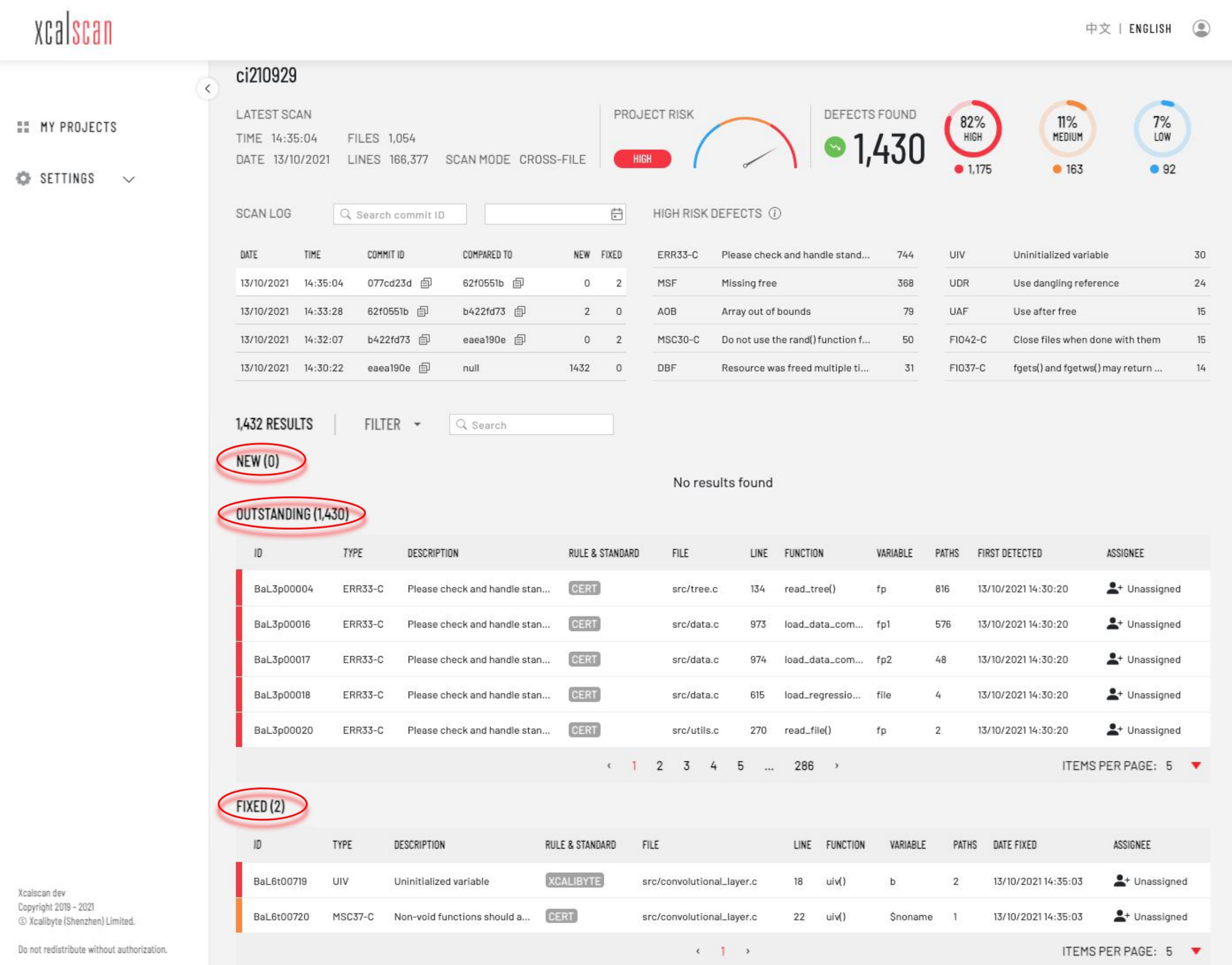
ICA:IssueChangeAnalysis

GB Mem. Required for ICA



Time (sec.) Spent for ICA





ci210929

LATEST SCAN
TIME 14:35:04 FILES 1,054
DATE 13/10/2021 LINES 166,377 SCAN MODE CROSS-FILE

PROJECT RISK HIGH

DEFECTS FOUND 1,430
82% HIGH 1,175
11% MEDIUM 163
7% LOW 92

SCAN LOG

DATE	TIME	COMMIT ID	COMPARED TO	NEW	FIXED
13/10/2021	14:35:04	077cd23d	62f0551b	0	2
13/10/2021	14:33:28	62f0551b	b422fd73	2	0
13/10/2021	14:32:07	b422fd73	eaea190e	0	2
13/10/2021	14:30:22	eaea190e	null	1432	0

HIGH RISK DEFECTS

ERR33-C	Please check and handle stand...	744	UIV	Uninitialized variable	30
MSF	Missing free	368	UDR	Use dangling reference	24
AOB	Array out of bounds	79	UAF	Use after free	15
MSC30-C	Do not use the rand() function f...	50	FIO42-C	Close files when done with them	15
DBF	Resource was freed multiple ti...	31	FIO37-C	fgets() and fgetws() may return ...	14

1,432 RESULTS | FILTER

NEW (0)

No results found

OUTSTANDING (1,430)

ID	TYPE	DESCRIPTION	RULE & STANDARD	FILE	LINE	FUNCTION	VARIABLE	PATHS	FIRST DETECTED	ASSIGNEE
BaL3p00004	ERR33-C	Please check and handle stan...	CERT	src/tree.c	134	read_tree()	fp	816	13/10/2021 14:30:20	Unassigned
BaL3p00016	ERR33-C	Please check and handle stan...	CERT	src/data.c	973	load_data_com...	fp1	576	13/10/2021 14:30:20	Unassigned
BaL3p00017	ERR33-C	Please check and handle stan...	CERT	src/data.c	974	load_data_com...	fp2	48	13/10/2021 14:30:20	Unassigned
BaL3p00018	ERR33-C	Please check and handle stan...	CERT	src/data.c	615	load_regressio...	file	4	13/10/2021 14:30:20	Unassigned
BaL3p00020	ERR33-C	Please check and handle stan...	CERT	src/utils.c	270	read_file()	fp	2	13/10/2021 14:30:20	Unassigned

‹ 1 2 3 4 5 ... 286 › ITEMS PER PAGE: 5

FIXED (2)

ID	TYPE	DESCRIPTION	RULE & STANDARD	FILE	LINE	FUNCTION	VARIABLE	PATHS	DATE FIXED	ASSIGNEE
BaL6t00719	UIV	Uninitialized variable	XCALIBYTE	src/convolutional_layer.c	18	uiv()	b	2	13/10/2021 14:35:03	Unassigned
BaL6t00720	MSC37-C	Non-void functions should a...	CERT	src/convolutional_layer.c	22	uiv()	$noname	1	13/10/2021 14:35:03	Unassigned

‹ 1 › ITEMS PER PAGE: 5

Xcalscan dev
Copyright 2019 - 2021
© Xcalibyte (Shenzhen) Limited.

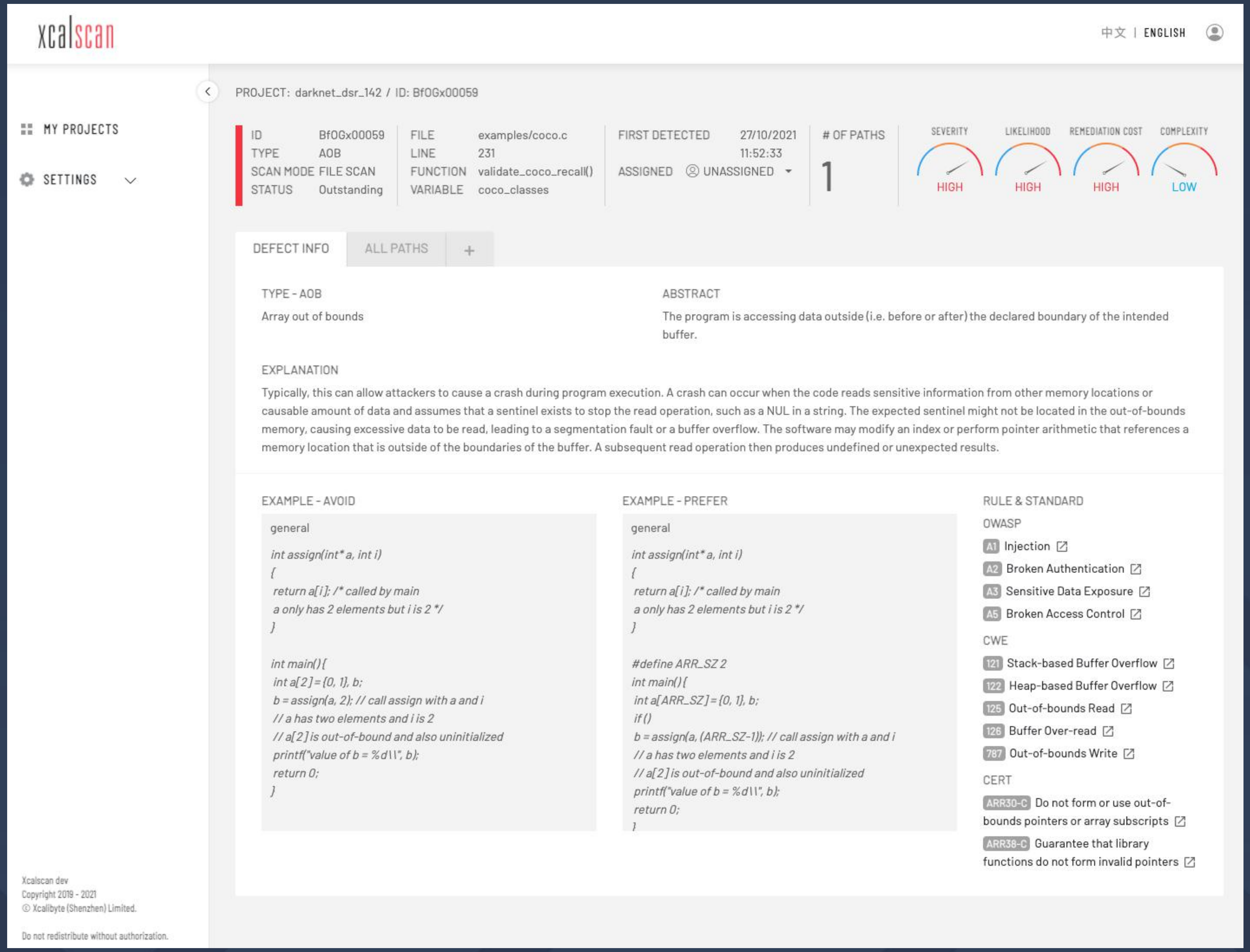
Do not redistribute without authorization.

- Rewrite Complex Algorithm with C++, Save Both Time and Space

PERFORMANCE IMPROVEMENTS

(cont)

- Remove Rule Information From RDB
  - ➢ Lazy Load of Rule Information
  - ➢ Reduced Memory Footprint During Query



PROJECT: darknet_dsr_142 / ID: BfOGx00059

ID	BfOGx00059	FILE	examples/coco.c	FIRST DETECTED	27/10/2021 11:52:33	# OF PATHS
TYPE	AOB	LINE	231	ASSIGNED	UNASSIGNED	1
SCAN MODE	FILE SCAN	FUNCTION	validate_coco_recall()			
STATUS	Outstanding	VARIABLE	coco_classes			

SEVERITY	LIKELIHOOD	REMEDIATION COST	COMPLEXITY
HIGH	HIGH	HIGH	LOW

TYPE - AOB

Array out of bounds

ABSTRACT

The program is accessing data outside (i.e. before or after) the declared boundary of the intended buffer.

EXPLANATION

Typically, this can allow attackers to cause a crash during program execution. A crash can occur when the code reads sensitive information from other memory locations or causable amount of data and assumes that a sentinel exists to stop the read operation, such as a NUL in a string. The expected sentinel might not be located in the out-of-bounds memory, causing excessive data to be read, leading to a segmentation fault or a buffer overflow. The software may modify an index or perform pointer arithmetic that references a memory location that is outside of the boundaries of the buffer. A subsequent read operation then produces undefined or unexpected results.

EXAMPLE - AVOID

```
general

int assign(int* a, int i)
{
 return a[i]; /* called by main
 a only has 2 elements but i is 2 */
}

int main(){
 int a[2]={0, 1}, b;
 b = assign(a, 2); // call assign with a and i
 // a has two elements and i is 2
 // a[2] is out-of-bound and also uninitialized
 printf("value of b = %d\\", b);
 return 0;
}
```

EXAMPLE - PREFER

```
general

int assign(int* a, int i)
{
 return a[i]; /* called by main
 a only has 2 elements but i is 2 */
}

#define ARR_SZ 2
int main(){
 int a[ARR_SZ]={0, 1}, b;
 if ()
 b = assign(a, (ARR_SZ-1)); // call assign with a and i
 // a has two elements and i is 2
 // a[2] is out-of-bound and also uninitialized
 printf("value of b = %d\\", b);
 return 0;
}
```

RULE & STANDARD

OWASP

- A1 Injection
- A2 Broken Authentication
- A3 Sensitive Data Exposure
- A5 Broken Access Control

CWE

- 121 Stack-based Buffer Overflow
- 122 Heap-based Buffer Overflow
- 125 Out-of-bounds Read
- 126 Buffer Over-read
- 787 Out-of-bounds Write

CERT

- ARR30-C Do not form or use out-of-bounds pointers or array subscripts
- ARR38-C Guarantee that library functions do not form invalid pointers

Xcalscan dev
Copyright 2019 - 2021
© Xcalibyte (Shenzhen) Limited.

Do not redistribute without authorization.

- Time to retrieve rule info (Read Only Data) – **>10x less**

DESIGN w/ FIRST PRINCIPLE

- Steps Involved
  - ➢ Collect Requirements and Facts
  - ➢ Partition Issues till Atomic
  - ➢ Reduce to Minimum: Discard Irrelevant & Unimportant
- What Mattered In The Refactor Effort:
  - ➢ There were Memory & I/O Bound Issues

QUANTITATIVE OBJECTIVES & MEASURE

- Performance Data Played a Key Role for Architecture Decision
  - ➢ Architecture Choices are Subjective
  - ➢ Measured Data are Objective
- Measurements on Data, Resource and Logic
  - ➢ Data Size – Footprint, Persistent
  - ➢ Data Processing Logic – Access Pattern Matters

QUANTITATIVE OBJECTIVES & MEASURE

- Strategies Applied in DB Refactor Project
  - ➢ Program Stability Issues Happened on Cases with Large Data Set
  - ➢ Systematically Measure Data Size, Memory Footprint
  - ➢ Track Response Time for APIs that is > 2 Seconds
- Policy Changed during and after DB Refactor Project
  - ➢ Performance Measured and Summarized Nightly

PROGRAM TO DETECT ISSUES INTELLIGENTLY

- Strategies Applied in DB Refactor Project
  - Log Data Characteristics at Building Block Level
  - Instrument Performance Monitor – White Box Approach
  - Nightly Regression Test Accelerates Project Integration
  - Use Option to Enable / Disable Functionality for Triage Purpose
  - Unified Log Format Facilitate Issue Reproduction

CODE WITH RIGHT PROGRAMMING LANGUAGE IN RIGHT PLACE

- Strategy Applied in DB Refactor Project
  - ➢ Easy DB manipulation, Java Stayed w/ Reduced Scope
  - ➢ Performance intensive, C++ : Pre-calculate Data, ICA function
  - ➢ Wrapper Functions, Added JavaScript for React Programming

REFACTOR OFTEN AND EARLY

- Recommend Book:
  - http://stepanovpapers.com/notes.pdf by Alex Stepanov
  - He invented Generic Programming and C++ STL
  - Decomposing An Application Into A Collection Of General–purpose Algorithms And Data Structures Makes It Robust
- Strategy Adopted: Incremental Phase–in Changes

RENEW KNOWLEDGE PERIODICALLY

- Bottom Line: You Own Your Own Career !
  - Get Ready for Your Next Job — In The Same Company Or Not
- Drive the Learning To Enhance Your Knowledge Framework
  - When And Where Will You Need New Techniques To Improve Your Project Architecture
  - Learning the Design Rationale Behind Open-Source Package
- Dedicate 8 Hours Per Week

严谨 – 架构师的必要特质

- 用科学方法来工作：实验尝试，小心求证
- 用 First Principle 来分析、思考
- 用性能数据来佐证、决策，并管控
- 用程序内置工具来监控架构劣坏
- 为算法适配程序语言
- 重构要及时并持续进行
- Stay Hungry, Stay Foolish

THANKS

Global
Architect Summit

AGENDA

- Case – Database Refactor (DR) Project
- Design with First Principle
- Define Quantitative Objectives and Measure Nightly
- Program to Detect Issues Intelligently
- Code with Right Programming Language in Right Place
- Refactor Often and Early
- Renew Knowledge Periodically
- Summary

IMPROVEMENTS AFTER DB REFACTOR

- DB Ingestion Time — **Now 4x faster**
- DB Query Time — **Now 20x faster**
- Response Time for UI dashboard (50 proj) — **2–3 sec, was 3 — 25 min**
- Enable Lage Project Scans
  - **File size 10x smaller; DB Ingestion Time 10x less; Smaller Memory Footprint, e.g. MySQL5.7 — 565MB vs. 63MB; 1807s vs. 43s; 22G vs. < 1G**
- Time and memory required to generate IssueChange
  - **Time 10x less; Memory 25x less, e.g. SQLite 20+min vs. 2 min; 10G vs. 0.4G.**
- Time to retrieve rule info (Read Only Data) – **>10x less**

SUMMARY

- 用科学方法来工作：实验尝试，小心求证
- 用First Principle 来思考：Connecting the Dot
- 用 OKR 来自我管理
- 用流水线来安排工作
- 在制高点审视全局
- 时时准备交班
- 时时刻刻提升自己的市场价值
- 每三年检讨自己在 Maslow's Hierarchy 的位置，步步为营
- 失败为成功之母，失之东隅，收之桑榆
- Stay Hungry, Stay Foolish