Jin Mingjian

Chinese Citizen, Ph.D., jin.phd@gmail.com

I am best known as tonystarkjin/ , a grand prizes hunter in the series of Alibaba Tianchi Engineering Competitions. The main prizes can be seen in the section of Awards.

This makes following records in the range of 2015 - 2020 Tianchi competition seasons:

- **Person with the Most Total Prizes** in the history of Alibaba Tianchi Engineering Competitions (2015 - 2020): Two silver cups, three medals, four awarded titles, numerous gifts
- **Person with the Most Total Money Rewards** in the history of Alibaba Tianchi Engineering Competitions (2015 - 2020)
- **Highest rewards in all solo participants (one-person teams)** in the history of Alibaba Tianchi Engineering Competitions (2015 - 2020), and for the two successive 2018 to 2019 seasons.

(It is better to check the wonderful in the section of Talks and Writings, not only the prizes. What I really want to get is not the rank#1, but **how joyful to explore kinds of engineering topics from an individual's effort**.)

Grand Prizes of Tianchi Engineering Competitions, 2015 - 2020

Awarded By Alibaba Group VP (Flink 2019)

Awarded By Manager of Alibaba Cloud Database Division (PolarDB 2018)

I am the founder of open-source high-performance foundationware, Landz. Landz, which is the first high-performance Java 8 ready stack in 2013, beats Netty 30% more in third-part benchmark with its own vectorizated http engine via original concise and composable library layers.

I am a bigdata expert. I have modified to the Storm sources to achieve 2x message throughput. I have re-implemented the Flink parquet file reader with native arrow based vectorized reader to reach 4x speed-up than presto-parquet reader and 2x

speed-up than Flink reader (in 1.19). And I have hacked the Cassandra client driver to get multi-core scale-up fairness for massive client writes. And futhermore, I have successfully constructed serveral dedicated realtime data clusters and warehouses with 10x cost-effective than traditonal open-source bigdata solutions (a.k.a., SQL-on-Hadoop, like Impala, Hive, HBase, Presto...). Yes, I am in finding the road to next-generation big data analysis.

I am contributing to kinds of open source systems and projects, including official PRs to Flink, Tensorflow(now LLVM) MLIR, Scala and Eclipse.

I am used to be as an expert in the Scala, Swift, Eclipse, Matlab.

I am a strong leader and have excellent communication skills.

Currently, I am interesting to reform scalable data intelligent infrastructures for massive data volume (from exabyte to zettabyte) in full spectrum, including the fields of storage design, query scheduling, query execution, query optimization, human - data interaction, under one modern distributed enviroment with fearless operation engineering.

I am also trying to mix some from data science into the data engineering.

Awards

- **Silver Medalist** of Apache Flink Geek Challenge Competition (2019)
- **Silver Medalist** of First Alibaba Cloud PolarDB Database Performance Competition (2018)
- **Winner Prize** of Third Alibaba Middleware Performance Challenge Competition (2017)
- **Three successive Week Stars** of Sina Weibo Prediction Challenge Competition (2015)
- **Google Summer of Code** 2010 Membership

Work History

- **Manager of Big Data Department**

  Tigerjoys, 2017.7 – 2020.4

  - Responsible for whole data department
  - Realtime Interactive Data Warehouse
    - Creative architecture
    - Peta-byte level data housing
    - Wide-dimension OLAP cube analysis, with 50+ dimensions mix-in query
    - Lighting-fast query, with 28-column and 6-table join-aggregation queries for billion wide rows done within 10 seconds
    - Extreme data ingestion throughput, with max 3.8 million wide rows writes per second observed
    - High concurrency, with max 8000+ queries per hours
    - 10x cost-effective compared to last generation platform, done with 7-node Dell commodity cluster

        - Highly stable, with zero downtime and zero production fault in its first two years
    - Data Middleware Platform
        - Data Input Abstraction and Government
            - dimensions and metrics can be customized
        - Data Output Abstraction and Government
            - formats can be choosed between csv, text-formated, json
        - Data Transportation Abstraction and Government
            - high concurrent connections supported
            - scale-out for multi-nodes
            - high availability for data query connections
    - Data visualization system for realtime bigdata
        - Interactive sql driven dashboard
    - File streaming based ETL tooling
        - Original and creative multi-threaded architecture
        - Works for any plain file stream
        - Saturated single hard-disk but designed to be scaled up to muliti-disks and scaled out to muliti-nodes
    - Heuristic balancing algorithm for data warehouse

- **Senior Architect/Director**

  AnG Tech Inc., 2015.10 - 2016.9

    - Responsible for whole data engineering team
    - Pluggable bidding algorithm framework for the DSP platform
    - **Unified Query Engine**
        - Unifies all heterogeneous query systems together
    - **DSP Insight**: lightning fast big data analysis platform for DSP in digital advertising
        - Ad-hoc SQL queries on ten-million-record data sets below 5 seconds, almost 50% query latency improvement than that of another 49-node pure Impala cluster on same data sets
        - support to write and read(scan) rather than other SQL on Hadoop, and gains 80%-500% faster than those of HBase
        - Extensions to Spark SQL and machine learning
        - Big Data driven personalized recommendation

- **Chief Platform Architect**

  Beijing KangYuan Inc.(Closed), 2014.5 - 2015.4

    - Responsible for whole platform engineering team
    - Hosting and showing the platform to kinds of investment teams(i.e., Intel China)
    - **Health Bigdata Platform for Wearable AIoT Devices**
        - Extreme single node performance based on state-of-the-art Landz stack
        - Zero-overhead encryption for AIoT transport(compared to the expensive HTTPS)
        - Load-balancing-friendly cluster designs with unlimited horizontal scaling high availability supported by the true async multicast of Landz
        - Hacking Cassandra client driver with Landz ’s dedicated async IO

pool for excellent throughput and latency
        - Cassandra based cloud storage.
        - ZStack/Cassandra/Kafka based mobile client message push system
        - Real-time AIoT sensor data analysis
        - Server/Rack/Datacenter bidding/hosting/planning/executing
        - Zero-downtime Devops
        - Data driven health recommendation algorithm
        - Geospatial interests query

- **Founder**

  Landz Project, 2013.10 - 2014.4

    - Machine-affinity Java 8 foundation with highest performance records in 2013
        - Self-host enabled with full stack completed
        - Carefully forged wait-free/lock-free zero-garbage zero-copy false-sharing-free concurrency facilities, which provides the only known correct bounded lock-free MPMC queue Java implementation in 2013
        - Pure Java lock-free offheap memory allocator, which is even faster than native JEMalloc
        - ZNR based clean-room crypto and other native functionalities implementation in Java. NO heartbleeding, NO Shellshock, NO Ghost in 2013
        - Self-balancing async thread pool with nanosecond level latency ITC(Inter Thread Communication) , 8x more efficient than Java’s fork-join pool in small tasks dispatching/balancing in 2013
        - New IO thread pattern to extract all possible cpu time from the Linux kernel with the epoll’s ET mode which is on par of io_uring's peformance only available in recent modern kernels
        - True async UDP/Multicast supporting with above excellent net io facilities
        - Zero-garbage vectorizated http engine is 3x times faster than Netty http parser, 5x faster than Node in 2013
        - Zero-garbage sync thread-safe logging API with minimum overhead, SSD saturated, 6x faster than the Log4j2 asynchronous logger in 2013
        - Composable, layered APIs provide the maximum engineering flexibility
        - 30% higher in the throughput than that of Netty by using the Techempower throughput benchmark test(plaintext) and very much better in latency in 2013
    - Incubated some Landz based start-up ideas

- **Senior Member of Technical Staff**

  Oracle Research and Development Center - Beijing,

  2010.7 - 2013.9

    - **Oracle WebLogic Server** Global Web Service Team
    - Driven new features:
        - Web service modularity and metro consolidation for WebLogic Server 12c
        - Web service client container support for Weblogic Server 12c

dwp(Development Web Profile, JavaEE6)
        - WebLogic web service's dynamic configuration and management
        - adapting WebLogic web service component passed with EE6 CTS
        - disabling WLS WSEE container(better than the corresponding of WebSphere)
        - mavenizing the build infra, implementing a new portable logging API for WebLogic Webservice
    - Numerous bug fixings for WebLogic web service component 11g(PS3/PS4/PS5) and 12c, JAXWS RI(Glassfish)/EJB Container/Servlet Container
    - Develop for maintaining the Java Specification Requests: JSR#224/JSR#109
    - Mentor new newbies in the team

Education

- **Doctor**, Institute of Electrical Engineering, Chinese Academy of Sciences, Electrical Engineering, September 2004 – 2010
- **Master**, Institute of Electrical Engineering, Chinese Academy of Sciences, Electrical Engineering, September 2001 – 2004
- **Bachelor**, Hua Zhong University of Science and Technology, Electrical Engineering, September 1997 – 2001

Talks and Writings

- Final Presentation to Apache Flink Geek Challenge Competition 2019 (English)
- Addendum to First Polardb Competition Series #1 - Impact of File System on Fast IO (Chinese)
- Final Presentation to First Alibaba Cloud PolarDB Database Performance Competition (English)