# Big Sensor-Generated Data Streaming Using Kafka and Impala for Data Storage in Wireless Sensor Network for $CO_2$ Monitoring

Rindra Wiska, Novian Habibie, Ari Wibisono, Widijanto Satyo Nugroho, and Petrus Mursanto
Faculty of Computer Science, Universitas Indonesia
Email : rindra.wiska@ui.ac.id, novian.habibie@ui.ac.id

***Abstract*—Wireless Sensor Network (WSN) is a system that have a capability to conduct data acquisition and monitoring in a wide sampling area for a long time. However, because of its big-scale monitoring, amount of data accumulated from WSN is very huge. Conventional database system may not be able to handle its big amount of data. To overcome that, big data approach is used for an alternative data storage system and data analysis process. This research developed a WSN system for $CO_2$ monitoring using Kafka and Impala to distribute a huge amount of data. Sensor nodes gather data and accumulated in temporary storage then streamed via Kafka platform to be stored into Impala database. System tested with data gathered from our-own made sensor nodes and give a good performance.**

## I. Introduction

As the world developed even more, the energy consumption around the world is increasing rapidly. Various form of energy used simultaneously in daily basis, such as fuel, electricity, food, etc. However, more consumed energy also means more waste produced as the remainder. Various waste have produced through the energy conversion process. But one matter that become the world's attention is Carbon Dioxide gas ($CO_2$).

The excessive amount of $CO_2$ in the air has become the worldwide issue because it can bring the bad effect to the environment in many aspects. It can cause several health issue such as nausea, breathing problem, vision problem, and so on[1]. It also can affect agricultural industry, even in the right amount $CO_2$ is useful for photosynthesis, in the excessive amount it can reduce the nutrients absorption rate for crops [2].

To encounter this problem, the monitoring system for $CO_2$ concentration is needed. It will be used to monitor the state of environment by conduct the mapping of $CO_2$ and make a prediction of its distribution in the future. To make the effective monitoring system in wide sampling area, the most suitable approach is by using Wireless Sensor Network (WSN).

WSN has an advantage in conduct the sampling and monitoring in the wide area. WSN use sampling device called sensor node which contains sensors as input, processor for process and aggregate the data, and communication module for data transmission. WSNs deployed in sampling area and gather data simultaneously. Then produced data will be gathered in single data center for further processing and analysis.

However, WSN system, especially which used to analyze and predict its data usually have one major issue in data collecting process : size of gathered data is huge. To make an analysis, small amount of monitoring data may not represent the actual condition of the sampling site. To overcome that, sampling duration needed is long. Moreover, to get a reliable data, the more amount of sampling site, the better quality of analysis is. Aggregated data from many sensor nodes in a long time of sampling duration will produce a huge size of data. The problem is, to process huge amount of data with conventional method will require a huge computational power and time, meanwhile data processing will not only occur once. There's always new data available from the sensor nodes, and new data may give an different effect to the analysis's result.

One of the most suitable solution to handle the huge data processing is by using Big Data approach. Big Data system is capable to store and process a huge amount of data in the most effective way. This approach can reduce data complexity and reduce its computation time in big amount of data.

This research focused on development of big data-tolerant system for WSN in $CO_2$ monitoring. Main purpose of this system is to provide a reliable and effective big data storage, processing, and analysis in the term of environment condition - focused on $CO_2$ concentration and distribution. This research focused on big data approach in Hadoop framework for embedded system sensor node.

## II. Wireless Sensor Network

Wireless Sensor Network (WSN) is a system that has an ability to conduct a data acquisition from a wide-range area using sensor nodes spread in the sampling area. Each sensor node works independently in acquiring data from its sampling point, and communicating each other forms a network between nodes to pass its acquired data to one point to be aggregated. WSN are widely used in many aspects. Its implementation varies from data gathering, surveillance, real-time monitoring, mapping, and so on [3].

One of the popular purposes of WSN is for environmental monitoring. Some related works conducted by various researches has succeeded gathered and analyzed many important data from environment with high performance and reliability. Al Turjman proposed a method to make a prolonged-lifetime WSN. It use Relay Nodes (RN) - a node specifically for data transmission - and Mobile RNs (MRN) - a RN that can be replaced to another point - as an addition to standard

sensor nodes. This system can maximizing its network lifetime meanwhile keeps its efficiency [4]. In another research, Jiang et. al. use WSN to monitor a behavior of honey bees related to environment state. It specifically gather a data of bees in their colony based on the condition of air temperature, humidity, and another environmental data [5].

## III. Big Data

There are so many definitions available to define what Big Data is. At least there are 43 different but similar definition of big data that can be found on [6] . One of the formal definition can be found on [7] : "an accumulation of data that is too large and complex for processing by traditional database management tools". Usually to simplify the definition, Big Data defined in 3 + 2 additional characteristics: Volume, Variety, Velocity[8] + Veracity and Value [9][10].

1) Volume, because it contains data with a huge size.
2) Variety, it contains so many variables in each instance.
3) Velocity, it produced in a quick time and almost continuous.
4) Veracity, data gathered may contains many errors.
5) Value, a goal that can be reached from the gathered data.

One of the capable framework To overcome the big amount of data is Hadoop. Hadoop is big data framework developed by Apache which is open source, fault tolerance, and scalable[11]. It developed to solve problem that use data in a huge amount-which is need certain technique in data storage and massive data processing. Hadoop is cheap but scalable because it can be built using common hardware setup but it an consists of many nodes that connected in one cluster.

Two main component of Hadoop are HDFS and MapReduce framework:

1) HDFS, a Hadoop Distributed File System: A storage system that distribute a large data among the node's storages. It distribute the data for three main advantages: 1)load balancing, 2)configurable block replication, and 3)recovery mechanism. HDFS have two layers: NameNode and DataNode. NameNode is a master-side that map the data into DataNode and main directory of a data contains property of a data such as name, permission, size, etc. DataNodes is a slave-side, provide a storage in the cluster. They serve the data and its permission of data access. In general, to acces data in HDFS, user need to access NameNode for file management.
2) MapReduce, a framework for parallel data processing. It will process data from HDFS in parallel. It contains two main process: Map Task and Reduce Task. Map Task generate a set of intermediate ¡key,value¿ pairs from data. Reduce Task will process the data to get final output. Two important module in MapReduce is JobTracker, which split task from user into multiple tasks, and TaskTracker, a subtask receiver that will run it on the nodes.

## IV. Big Data Approach in Wireless Sensor Network

As WSN have many nodes and deployed in a wide area, WSN fulfill all characteristic of the system that needs big data approach. It fulfill all big data characteristics in every aspects:

1) Volume
   WSN use many nodes for sampling for a certain time, so amount of gathered data is depend to how wide the sampling area is. Usually WSN covered a wide area and conduct the sampling for a long time e.g months, even years.
2) Variety
   To have a strong conclusion, data gathered must have several parameters that represent the actual condition of the sampling site. The problem is, sometimes the correlation between parameters and which parameter have a biggest contribution for an inference process is unknown. The best way to overcome this is to use as much as possible parameters available and conduct correlation analysis test between parameters later after data has gathered.
3) Velocity
   To give the highest accuracy, interval between sampling must be minimal. Because bigger the intervals, important condition or state between interval may be not recorded.
4) Veracity
   When using a sensor node, errors in data is common. Placed in a remote area and unattended, sensor node may risked by hardware or software failure. High uncertainty in sampling area also give an extra noise to gathered data.
5) Value
   Bigger the data size, bigger its value to inference process. By using more data, we can get more accurate and reliable results which can decrease the effect of noised and outliers data with the overwhelming amount of data.

Big data approach also started to implemented in various wide-range monitoring using WSN and combined with Internet of Things (IoT) concept. Progressively, people started to depend more on data than the infrastructure itself. Many IoT implementation, beside it have its own specific service to the users, now it can be enhanced to the wider contribution by only with use its distributed user data in its networks.

Some related works already combining WSN with big data concept for wide-range monitoring, especially in environmental monitoring. Zheng et. al. develop an air monitoring system with big data approach for data inference process. It use sensory streaming data from various parameters (meteorology data, traffic condition land uses, etc.) to infer the concentration of PM2.5 in Beijing using neural network and conditional random field (CRF) with high accuracy and fast processing time - near real-time[12]. In another research, Ong et. al. use a time series data of PM2.5 added with another meteorological data from 52 cities in Japan over a two year period. It analyzed with Deep Recurrent Neural Network (DRNN)

to produce a time-series prediction of PM2.5 for 12 hours ahead[13]. Hasenfratz et. al. use 50 million data of ultrafine particles (UFP) measurements from mobile sensor nodes of two years added with additional data e.g. traffic data to built a high-resolution spatio-temporal pollution maps of Zurich. It analyzed using Generalized Adaptive Models (GAMs) to develop land-use regression (LUR) model[14].

## V. Methodology

This research developed a WSN system integrated with big data approach. For data input, this system use sensor nodes as data gathering device which deployed in sampling area. Sensor nodes will gather data and aggregate it in one centralized temporary storage. In the next step, temporary storage that acts as producer, will transmit aggregated data into server using kafka. The server acts as consumer wil get the data and then transmit the data into Impala database. Methodology diagram of the system can be seen on figure 2.

## VI. System Design

As a WSN system, proposed system still use standard scheme of WSN as can be seen on figure **??**. Its major change occurred on its data storage system. While standard WSN only use one data storage, the proposed system use multiple data storages which its data distributed between them using Hadoop Distributed File System (HDFS). Proposed scheme of the system can be seen on figure 1

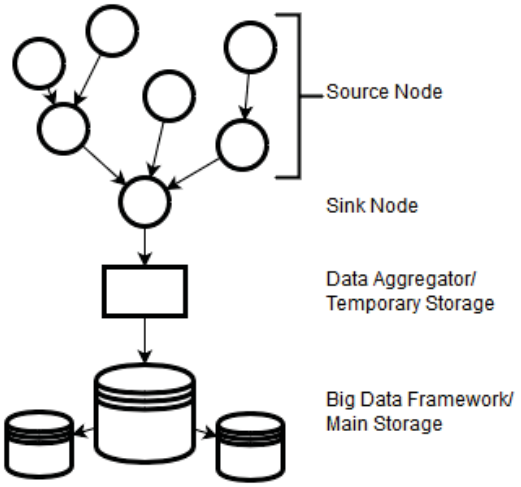


Fig. 1: General architecture of Wireless Sensor Network with Big Data approach

The system built by two main component : sensor node for $CO_2$ monitoring, and big data storage system.

### A. Sensor Node Implementation

This research use a sensor node for monitoring $CO_2$ concentration in the air included another supportive parameters : air humidity, air temperature, and light intensity of the sampling site. This sensor node built using Arduino board combined with sensor modules, timer module, and radio module. Details

TABLE I: My caption

| Components | Details |
|---|---|
| **Platform** | Arduino/Genuino Mega 2560 |
| **Microcontroller** | ATMega2560 |
| **Input Voltage** | 12V |
| **Sensors** | $CO_2$ Sensor COZIR Ambient 10K + RH/T (Humidity & Temperature) |
| | Light Sensor Module LM393 |
| **Timer** | RTC Module DS3231 |
| **Communication** | Radio Module NRF24L01+ |

of the hardware used in sensor node can be seen on table **??** ,architecture of sensor node can be seen on figure 3, and an implementation of sensor node can be seen on figure 4. For communication, sensor nodes communicating each other using static routing scheme. Sensor node acquire data form environment every five second and its result transmitted into temporary storage to be accumulated. Temporary storage used in this research is Single Board Computer (SBC) Raspberry Pi 2 provided with internet connection to aggregate data from sensors and send it into HDFS to be analyzed.

### B. Big Data Infrastructure

This research used Kafka for ditributed messages and used Impala database for data storage.

*1) Kafka:* Kafka is a framework for distributed systems that was developed to collect and distribute a massive message[15]. The basic architecture of Kafka consist of producers, brokers and consumers[16]. To maintain the messages kafka using topics. The topics will be writen by producers to ditribute data to consumers. In this research, raspberry pi 2 act as producer to send the collect data to consumers through a broker in server.

*2) Impala:* Impala is a SQL query engine that designed spesiffically to leverage the flexibility and scalability of Hadoop [17]. The beta release of Impala was in October 2012. Impala is built using C++ and Java. To reduce latency, Impala using architecture based on daemon processes. The daemon processes are responsible for doing query execution on Hadoop infrastructure. Xiaopeng Li and Wenli Zhou compared the performance of Impala with Hive and Spark SQL. The result of the comparison is Impala has has the fastest query speed [18].

## VII. System Performance

To measure system performance, this experiment use three Raspberry Pi which act as producers. The producers send data with same topic to consumer as shown in figure 5. Data used in this experiment is the original data obtained from $CO_2$ sensor node. The $CO_2$ sensor nodes collect data every three seconds.

The purpose of this experiments is to calculate the latency of data transmission by the producer. The experiments run for 420 minutes and figure 6 is the average of latency time per minute. The average time of each producer is: producer 1 with an average latency of 1.73 ms, producer 2 with an average latency 13.21 ms, and producer 3 with an average latency 13.19 ms. Latency time of producer 1 exceeds the

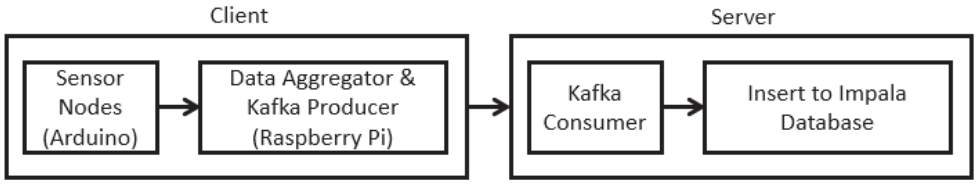


Fig. 2: Methodology of the system

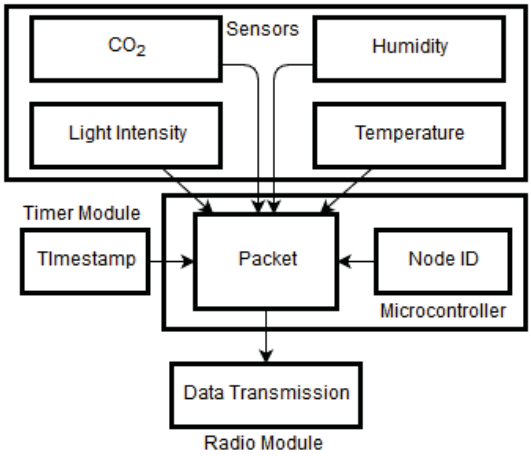


Fig. 3: Architecture of sensor node used in the system

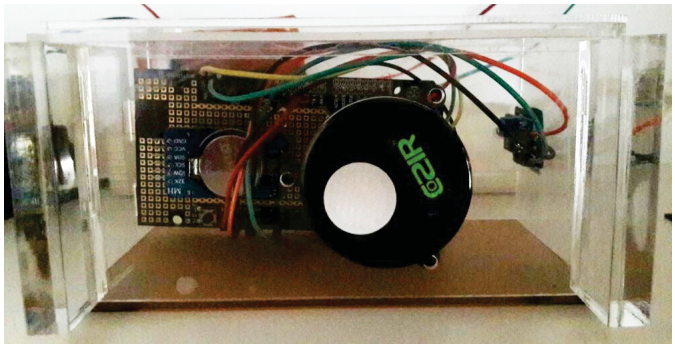

Fig. 4: Implementation of sensor node for the system

latency time of other producers because of producer 1 is at the same network with the kafka server.

## VIII. Conclusion

Wireless Sensor Network (WSN) is a system that produces big amount of data. Because of that, the system that capable to handle the big amount of data is needed. To overcome that, this research integrating standard WSN system with big data framework with streaming data approach. In implementation, the system used our own-made sensor nodes which gather $CO_2$ data alongside with another environmental data : air temperature, air humidity, and light intensity. Using Kafka and Impala database, this system capable to transmit huge amount

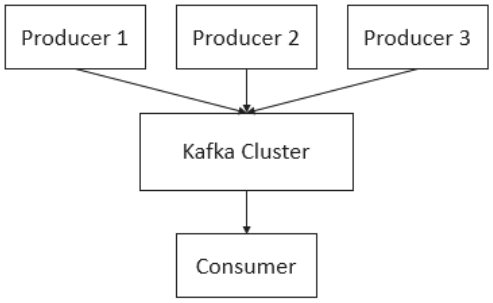


Fig. 5: System Implementation

of data in near real-time. Overall, proposed system give a good performance and reliable output.

## IX. Future Development

This research only focused on developing Wireless Sensor Network (WSN) with big data infrastructure. For future development, this system can be improved in the term of data analysis for prediction and data obtained can also be used to distribute the deployment of $CO_2$ in environment such as research on odor [19].

## X. Acknowledgment

This work was supported by Indonesian Directorate General of Higher Education (DIKTI) funding in 2015, namely Penelitian Unggulan Perguruan Tinggi (PUPT) with the grant number is 1070/H2.R12/HKP.05.00/2016.

## References

[1] BLM, "Health risk evaluation for carbon dioxide (co 2 )." [Online]. Available: http://www.blm.gov/style/medialib/blm/wy/information/NEPA/cfodocs/howell.Par.2800.File.dat/25apxC.pdf

[2] L. H. Dietterich and et al, "Impacts of elevated atmospheric co2 on nutrient content of important food crops," *Scientific Data*, vol. 2, no. 150036, 2015. [Online]. Available: http://www.nature.com/articles/sdata201536

[3] B. Rashid and M. H. Rehmani, "Applications of wireless sensor networks for urban areas: A survey," *Journal of Network and Computer Applications*, vol. 60, pp. 192 – 219, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804515002702

[4] F. M. Al-Turjman, H. S. Hassanein, and M. Ibnkahla, "Towards prolonged lifetime for deployed {WSNs} in outdoor environment monitoring," *Ad Hoc Networks*, vol. 24, Part A, pp. 172 – 185, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570870514001905

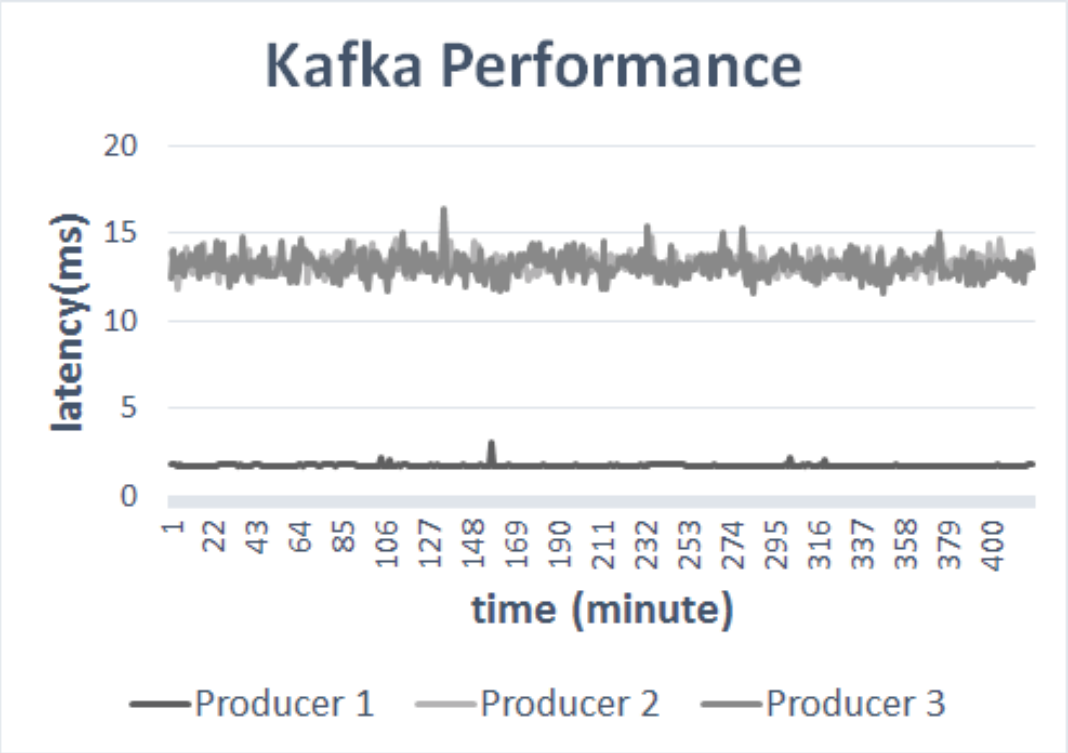


Fig. 6: Kafka Performance

[5] J.-A. Jiang, C.-H. Wang, C.-H. Chen, M.-S. Liao, Y.-L. Su, W.-S. Chen, C.-P. Huang, E.-C. Yang, and C.-L. Chuang, "A wsn-based automatic monitoring system for the foraging behavior of honey bees and environmental factors of beehives," *Computers and Electronics in Agriculture*, vol. 123, pp. 304 – 318, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169916300709

[6] datascience@berkeley, "What is big data?" September 2014. [Online]. Available: https://datascience.berkeley.edu/what-is-big-data/

[7] Merriam-Webster, "Definition of big data," September 2016. [Online]. Available: http://www.merriam-webster.com/dictionary/big data

[8] R. Price, "Volume, velocity and variety: Key challenges for mining large volumes of multimedia information," in *Proceedings of the 7th Australasian Data Mining Conference - Volume 87*, ser. AusDM '08. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2008, pp. 17–17. [Online]. Available: http://dl.acm.org/citation.cfm?id=2449288.2449292

[9] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Research*, vol. 2, no. 1, pp. 2 – 11, 2015, special Issue on Computation, Business, and Health Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214579615000118

[10] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11036-013-0489-0

[11] Apache, "What is apache hadoop." [Online]. Available: http://hadoop.apache.org/

[12] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '13. New York, NY, USA: ACM, 2013, pp. 1436–1444. [Online]. Available: http://doi.acm.org/10.1145/2487575.2488188

[13] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamic pre-training of deep recurrent neural networks for predicting environmental monitoring data," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 760–765.

[14] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, Part B, pp. 268 – 285, 2015, selected Papers from the Twelfth Annual {IEEE} International Conference on Pervasive Computing and Communications (PerCom 2014). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119214001928

[15] Kafka, "Kafka is a distributed streaming platform. what exactly does that mean?" [Online]. Available: https://kafka.apache.org/intro

[16] Z. Wang, W. Dai, F. Wang, H. Deng, S. Wei, X. Zhang, and B. Liang, "Kafka and its using in high-throughput and reliable message distribution," in *2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*, Nov 2015, pp. 117–120.

[17] M. Kornacker, A. Behm, V. Bittorf, T. Bobrovytsky, C. Ching, A. Choi, J. Erickson, M. Grund, D. Hecht, M. Jacobs, I. Joshi, L. Kuff, D. Kumar, A. Leblang, N. Li, I. Pandis, H. Robinson, D. Rorke, S. Rus, J. Russell, D. Tsirogiannis, S. Wanderman-Milne, and M. Yoder, "Impala: A modern, open-source sql engine for hadoop." [Online]. Available: http://impala.io/

[18] X. Li and W. Zhou, "Performance comparison of hive, impala and spark sql," in *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 1, Aug 2015, pp. 418–423.

[19] W. Jatmiko, Y. Ikemoto, T. Matsuno, T. Fukuda, and K. Sekiyama, "Distributed odor source localization in dynamic environment," in *IEEE Sensors, 2005.*, Oct 2005, pp. 4 pp.–.